

# An Extension of Generalized Linear Models to Finite Mixture Outcome Distributions

Andrew M. Raim\*, Nagaraj K. Neerchal<sup>†</sup> & Jorge G. Morel<sup>†</sup>

\*Center for Statistical Research and Methodology, U.S. Census Bureau

<sup>†</sup>Department of Mathematics and Statistics, University of Maryland, Baltimore County

## Abstract

Finite mixture distributions arise in sampling a heterogeneous population. Data drawn from such a population will exhibit extra variability relative to any single subpopulation. Statistical models based on finite mixtures can assist in the analysis of categorical and count outcomes when standard generalized linear models (GLMs) cannot adequately account for variability observed in the data. We propose an extension of GLM where the response is assumed to follow a finite mixture distribution, while the regression of interest is linked to the mixture's mean. This approach may be preferred over a finite mixture of regressions when the population mean is the quantity of interest; here, only a single regression function must be specified and interpreted in the analysis. A technical challenge is that the mean of a finite mixture is a composite parameter which does not appear explicitly in the density. The proposed model is completely likelihood-based and maintains the link to the regression through a certain random effects structure. We consider typical GLM cases where means are either real-valued, constrained to be positive, or constrained to be on the unit interval. The resulting model is applied to two example datasets through a Bayesian analysis: one with success/failure outcomes and one with count outcomes. Supporting the extra variation is seen to improve residual plots and to appropriately widen prediction intervals.

## 1 Introduction

The Generalized Linear Model (GLM) is heavily used by researchers and practitioners for regression analysis on categorical, count, and continuous outcomes (McCullagh and Nelder, 1989). Standard GLM theory assumes an exponential family distribution, such as Poisson to model counts and Binomial to model success/failure data. These distributions are limited in the amount of variability they can express. GLM users often encounter the issue of overdispersion, where the data exhibit variability which cannot be expressed by the model. This can manifest itself in a number of ways, depending on the specific nature of the overdispersion and its departure from the model. For example, assuming independence in clustered data can result in standard error estimates which are too small and lead to tests with an inflated type I error rate (Morel and Neerchal, 2012, Chapter 1).

The objective of this paper is to extend the GLM so that a finite mixture of  $J$  simpler densities can be used as the distribution for the response. There is a well-established literature on finite mixtures of regressions,

---

This paper is released to inform interested parties of ongoing research and to encourage discussion of work in progress. Any views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.

For correspondence:

A.M. Raim ([andrew.raim@gmail.com](mailto:andrew.raim@gmail.com))

Center for Statistical Research and Methodology

U.S. Census Bureau

Washington, D.C. 20233, U.S.A.

in which each component distribution of a finite mixture is linked to a separate regression (Frühwirth-Schnatter, 2006). An analyst may employ a finite mixture of regressions model if heterogeneity is suspected in the relationship between covariate  $\mathbf{x}$  and response  $y$  among sampled units, yet not enough is known to model the heterogeneity explicitly. Specifying regressions for  $J$  latent subpopulations may complicate model selection in practice. Often, the interest may be in modeling the mean response, and heterogeneity is simply a nuisance rather than a target for inference. This motivates us to formulate the Mixture Link model, which uses a finite mixture to capture extra variation, but constrains the mean of the finite mixture to be linked to a single regression function. The mean of a finite mixture is composed of multiple parameters which may not appear directly in the likelihood. Central to the development of Mixture Link is the set in which the link constraint is honored. In the case of positive-valued means, this constraint set is a polytope, while for probability-valued means it is the intersection of a polyhedron and a unit cube. For real-valued means, the constraint set is the basis of a linear space. A random effects structure is assumed on this set to complete specification of the likelihood. Under Poisson and Normal outcome types, the random effects can be integrated out to yield a tractable form for the density. The case of Binomial outcomes is more computationally challenging. Taking a Bayesian approach to inference, a simple Random-Walk Metropolis-Hastings sampler can be used for the Normal and Poisson Mixture Link models. For Binomial outcomes, we consider a Metropolis-within-Gibbs sampler with data augmentation to avoid repeated evaluation of the marginal density.

A number of methods have been established to handle overdispersion. Morel and Neerchal (2012) provide an overview in the settings of count and categorical data. One common approach is to extend a basic distribution by assuming the presence of latent random variables, and then integrating them out. The Beta-Binomial (Otake and Prentice, 1984), Zero-Inflated Binomial (Hall, 2000), and Random-Clumped Binomial (Morel and Nagaraj, 1993) distributions are all obtained in this way starting from the Binomial distribution. Similarly, the negative Binomial and zero-inflated negative Binomial distributions (Hilbe, 2011) are obtained starting from the Poisson distribution. In this same way, the t-distribution (Liu and Rubin, 1995) may be considered an overdispersion model relative to the normal distribution. Generalized Linear Mixed Models are obtained by adding random effects to the regression function (McCulloch et al., 2008); the marginal likelihood of the outcomes usually cannot be written without an integral for non-normal outcomes. Quasi-likelihood methods extend the likelihood in ways that do not yield a proper likelihood, but allow inference to be made on regression coefficients. A simple quasi-likelihood is obtained from placing a dispersion multiplier to the variance (Agresti, 2002, Section 4.7). The method of Wedderburn (1974) requires specification of only the mean-variance relationship to form a system of equations and carry out inference. Generalized Estimating Equations (GEE) is a quasi-likelihood method for grouped data where the analyst assumes a working correlation structure for observations taken within a subject (Hardin and Hilbe, 2012). Some Bayesian overdispersion methods are discussed in the collection assembled by Dey et al. (2000); for example, Basu and Mukhopadhyay (2000) consider generalizing the link function of a GLM to a mixture distribution and Dey and Ravishanker (2000) propose generalized exponential families for the outcome. More recently, Klein et al. (2015) proposed a Bayesian approach to generalized additive models under the Zero-Inflated Negative Binomial model to estimate complicated regression functions.

The rest of the paper proceeds as follows. Section 2 formulates the Mixture Link general model. Section 3 develops Mixture Link under probability-valued means, with special attention given to Binomial outcomes. Sections 4 and 5 develop Mixture Link for positive- and real-valued means, respectively, and obtain specific models for Poisson and Normal outcomes. Section 6 presents example data analyses with Mixture Link Binomial and Mixture Link Poisson. Finally, Section 7 concludes the paper. The `mixlink` package for R (available from <http://cran.r-project.org>) provides much of the Mixture Link functionality discussed in this paper.

## 2 Mixture Link Formulation

The usual GLM formulation is based on a density in the exponential dispersion family,

$$f(y \mid \theta, \phi) = \exp \left\{ \frac{\theta y - b(\theta)}{a(\phi)} + c(y; \phi) \right\}, \quad (2.1)$$

where  $\theta$  is the canonical parameter which influences the mean and  $\phi$  is the dispersion parameter. Here it can be shown that  $E(y) = b'(\theta)$  and  $\text{Var}(y) = a(\phi)b''(\theta)$ , and expressions for the score vector and information matrix can be obtained (Agresti, 2002, Section 4.4). Estimation can be carried out routinely, using Newton-Raphson or scoring algorithms to compute maximum likelihood estimates, or standard MCMC algorithms for a Bayesian analysis. Our objective is to modify this framework to allow a finite mixture as the outcome distribution, establishing a link between the mixture mean and a regression function of interest. Because finite mixtures can support more variation than distributions of the form (2.1), this extension should naturally support variation beyond standard GLMs. We are especially interested in finite mixtures of three common GLM outcome types: Normal, Binomial, and Poisson.

Consider a random variable  $Y$  following the finite mixture distribution,

$$f(y \mid \boldsymbol{\theta}) = \sum_{j=1}^J \pi_j g(y \mid \boldsymbol{\theta}_j). \quad (2.2)$$

Here, the mixing proportions  $\boldsymbol{\pi} = (\pi_1, \dots, \pi_J)$  belong to the probability simplex  $\mathcal{S}^J = \{\boldsymbol{\lambda} \in [0, 1]^J : \lambda_j \geq 0, \boldsymbol{\lambda}^T \mathbf{1} = 1\}$ . The densities  $g(y \mid \boldsymbol{\theta}_j)$  belong to a common family parameterized by  $\boldsymbol{\theta}_j = (\mu_j, \boldsymbol{\phi}_j)$ , consisting of a mean parameter  $\mu_j = \int y g(y \mid \boldsymbol{\theta}_j) d\nu(y)$  and where all other parameters are contained in  $\boldsymbol{\phi}_j$ . Writing  $\nu$  as the dominating measure for densities  $g$  allows expectations over discrete and continuous random variables to be treated with a common integral notation. The overall expected value is  $E(Y) = \sum_{j=1}^J \pi_j \mu_j = \boldsymbol{\pi}^T \boldsymbol{\mu}$ . The  $\mu_j$  may naturally be restricted to a subset of  $\mathbb{R}$ , depending on the outcome type. For example, if  $Y$  is a count,  $\mu_j \in [0, \infty)$  often represents a rate. Alternatively, if  $Y$  is the number of successes among  $m$  trials, which result in either success or failure, then  $\mu_j \in [0, 1]$  can represent the probability of a success. In general, denote the natural space of  $\mu_j$  as  $\mathcal{M}$ , so that  $\boldsymbol{\mu} = (\mu_1, \dots, \mu_J)$  is an element of  $\mathcal{M}^J$ .

In a regression setting, we observe a random sample  $Y_1, \dots, Y_n$  from the finite mixture

$$f(y_i \mid \boldsymbol{\theta}_i) = \sum_{j=1}^J \pi_j g(y \mid \mu_{ij}, \boldsymbol{\phi}_{ij}), \quad (2.3)$$

with an associated (fixed) predictor  $\mathbf{x}_i \in \mathbb{R}^d$ , for  $i \in \{1, \dots, n\}$ . As in the traditional GLM, we wish to link  $E(Y_i)$  to a regression function such as  $\mathbf{x}_i^T \boldsymbol{\beta}$  through an inverse link function  $G$ . To simplify expressions in the rest of the paper, denote  $\vartheta(\mathbf{x})$  as the inverse-linked regression  $G(\mathbf{x}^T \boldsymbol{\beta})$ . We will write  $\vartheta_i = G(\mathbf{x}_i^T \boldsymbol{\beta})$  for brevity when specifically referring to the  $i$ th observation, and  $\vartheta$  in place of  $\vartheta(\mathbf{x})$  when not emphasizing a specific observation. With this notation, our objective is to link

$$\boldsymbol{\pi}^T \boldsymbol{\mu} = \vartheta_i. \quad (2.4)$$

The left-hand side of (2.4) must vary with the observation for the link to be achievable. In this work, we will assume that subpopulation means  $\boldsymbol{\mu}_i = (\mu_{i1}, \dots, \mu_{iJ})$  are specific to the  $i$ th observation, but that mixing proportions  $\boldsymbol{\pi}$  are common across observations. In contrast to the traditional GLM setting,  $\boldsymbol{\pi}^T \boldsymbol{\mu}_i$  is a composite parameter which does not appear directly in the density of  $Y_i$ . Therefore, we cannot simply plug  $\vartheta_i$  into the likelihood.

To enforce (2.4), consider the set

$$A(\vartheta, \boldsymbol{\pi}) = \{\boldsymbol{\mu} \in \mathcal{M}^J : \boldsymbol{\mu}^T \boldsymbol{\pi} = \vartheta\}. \quad (2.5)$$

For a given  $\boldsymbol{\beta}$  and  $\boldsymbol{\pi}$ , restricting ourselves to  $\boldsymbol{\mu}_i \in A(\vartheta_i, \boldsymbol{\pi})$  is equivalent to enforcing the link. We will write  $A$  as a shorthand for  $A(\vartheta, \boldsymbol{\pi})$  and  $A_i$  for  $A(\vartheta_i, \boldsymbol{\pi})$ . Our approach will be to take  $\boldsymbol{\mu}_i$  as a random effect

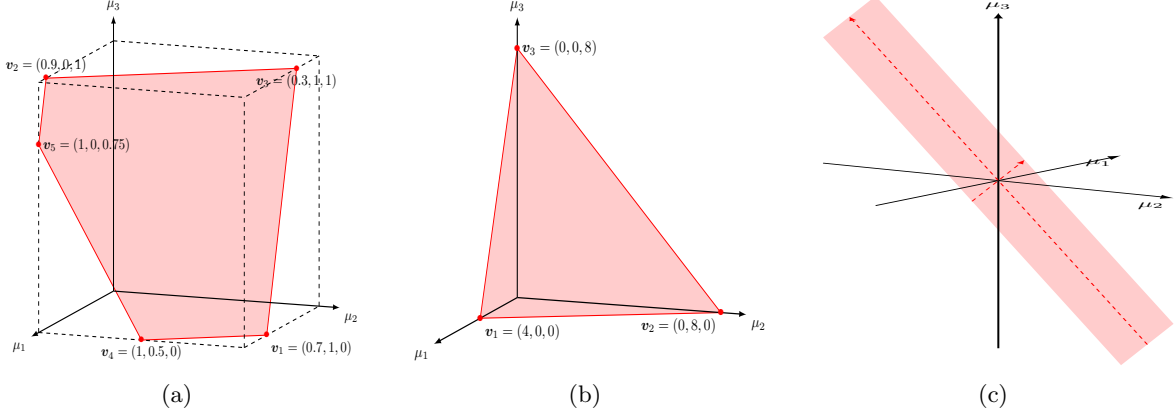


Figure 1: Examples of the set  $A(\vartheta, \boldsymbol{\pi})$  in dimension  $J = 3$ : (a) probability-valued means with  $\boldsymbol{\pi} = (0.5, 0.3, 0.2)$  and  $\vartheta = 0.65$ , (b) positive means with  $\boldsymbol{\pi} = (0.5, 0.25, 0.25)$  and  $\vartheta = 2$ , (c) real-valued means with  $\boldsymbol{\pi} = (0.5, 0.3, 0.2)$  and  $\vartheta = 0$ .

drawn from set  $A(\vartheta_i, \boldsymbol{\pi})$ . In Sections 3, 4, and 5 we will consider several commonly used choices of the space  $\mathcal{M}$ —the unit interval, the positive real line, and the real line respectively—to determine an appropriate distribution for  $\boldsymbol{\mu}_i$ . Figure 1 displays an example of the set  $A(\vartheta_i, \boldsymbol{\pi})$  for each of these three cases. [Boyd and Vandenberghe \(2004\)](#) is a useful reference for basic concepts in the analysis of convex sets which emerge in the remainder of the paper. Note that  $\mathbf{x}_i = 1$  may be taken for all  $i = 1, \dots, n$  to yield a non-regression version of Mixture Link.

Selection of a distribution over  $A(\vartheta, \boldsymbol{\pi})$  determines the density of  $Y_i$ ,

$$\begin{aligned} f(y_i \mid \boldsymbol{\beta}, \boldsymbol{\pi}, \boldsymbol{\phi}_i) &= \int \sum_{j=1}^J \pi_j g(y_i \mid \mu_{ij}, \boldsymbol{\phi}_{ij}) \cdot f_{A^{(i)}}(\boldsymbol{\mu}_i) d\boldsymbol{\mu}_i \\ &= \sum_{j=1}^J \pi_j \int g(y_i \mid w, \boldsymbol{\phi}_{ij}) \cdot f_{A_j^{(i)}}(w) dw. \end{aligned} \quad (2.6)$$

Here,  $f_{A^{(i)}}$  represents the  $J$ -dimensional random effects density over  $A(\vartheta_i, \boldsymbol{\pi})$  and  $f_{A_j^{(i)}}$  represents the marginal density of the  $j$ th coordinate. In the trivial case  $J = 1$ , there is only a single point in  $A(\vartheta_i, \boldsymbol{\pi})$ , and  $f(y_i \mid \boldsymbol{\beta}, \boldsymbol{\pi}, \boldsymbol{\phi}_i)$  simplifies to  $g(y_i \mid \vartheta_i, \boldsymbol{\phi}_{i1})$ . In general, evaluating  $f(y_i \mid \boldsymbol{\beta}, \boldsymbol{\pi}, \boldsymbol{\phi}_i)$  requires computation of  $J$  univariate integrals, which can be achieved numerically using quadrature or other standard techniques. This can become a computational burden if  $f(y_i \mid \boldsymbol{\beta}, \boldsymbol{\pi}, \boldsymbol{\phi}_i)$  must be computed many times (e.g. for a simulation or iterative estimation procedure) or if  $f_{A_j^{(i)}}(w)$  is difficult to evaluate. By construction,  $E(Y_i) = \vartheta_i$ , but variance and other moments depend on  $g$  and the distribution of  $\boldsymbol{\mu}_i$ . As in more basic finite mixture models, the value of density (2.6) is invariant to permutations of the subpopulation labels  $\{1, \dots, J\}$ .

### 3 Probability-Valued Means

Consider the setting  $\mathcal{M} = [0, 1]$ , which is useful for Bernoulli or Binomial data where means represent probabilities. It is straightforward to verify that  $A(\vartheta_i, \boldsymbol{\pi}) = \{\boldsymbol{\mu} \in [0, 1]^J : \boldsymbol{\mu}^T \boldsymbol{\pi} = \vartheta_i\}$  is a bounded convex set in  $\mathbb{R}^J$ . Therefore, we have the decomposition

$$A(\vartheta_i, \boldsymbol{\pi}) = \left\{ \sum_{\ell=1}^{k_i} \lambda_\ell \mathbf{v}_\ell^{(i)} : \boldsymbol{\lambda} \in \mathcal{S}^{k_i} \right\} = \left\{ \mathbf{V}^{(i)} \boldsymbol{\lambda} : \boldsymbol{\lambda} \in \mathcal{S}^{k_i} \right\}. \quad (3.1)$$

The  $J \times k_i$  matrix  $\mathbf{V}^{(i)}$  is composed of the columns  $\mathbf{v}_1^{(i)}, \dots, \mathbf{v}_{k_i}^{(i)}$  which are vertices of  $A(\vartheta_i, \boldsymbol{\pi})$ . Any element  $\boldsymbol{\mu} \in A(\vartheta_i, \boldsymbol{\pi})$  can be written as a convex combination of these vertices. The matrix  $\mathbf{V}^{(i)}$  depends on both  $\boldsymbol{\pi}$  and  $\vartheta_i$ ; both its elements and the dimension  $k_i$  may vary with the observation  $i = 1, \dots, n$ . The vector  $\boldsymbol{\lambda}^{(i)}$  belongs to the probability simplex  $\mathcal{S}^k$ .

The Minkowski-Weyl decomposition of a polyhedron is  $P = \{\sum_{\ell=1}^k \lambda_\ell \mathbf{v}_\ell : \boldsymbol{\lambda} \in \mathcal{S}^k\} + \{\sum_{\ell=1}^h \lambda_\ell \boldsymbol{\xi}_\ell : \boldsymbol{\lambda} \geq 0\}$ , relative to extreme points  $\mathbf{v}_1, \dots, \mathbf{v}_k$  (i.e. vertices) and extreme directions  $\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_h$  of  $P$ . The set  $A_i$  in (3.1) is a polytope, a bounded polyhedron not having extreme directions, for which we need only consider extreme points. Assuming a distribution on the coefficients of the Minkowski-Weyl decomposition has been advocated by Danaher et al. (2012), who sought a class of priors to enforce biologically motivated polyhedral constraints in a Bayesian analysis.

A natural choice for a random effects distribution on  $\mathcal{S}^{k_i}$  is  $\boldsymbol{\lambda}^{(i)} \stackrel{\text{ind}}{\sim} \text{Dirichlet}_{k_i}(\boldsymbol{\alpha})$ . However, this choice leads to each component of  $\boldsymbol{\mu}_i = \mathbf{V}^{(i)} \boldsymbol{\lambda}^{(i)}$  following the distribution of a linear combination of a  $k$ -dimensional Dirichlet. This distribution is computationally impractical; for example, its density has no known closed form for general  $k$  (Provost and Cheong, 2000). Our approach will first be to state the model using a Dirichlet random effect, then to state a more practical form of the model using Beta random effects with matched first and second moments. This ensures, for example, that  $\mathbb{E}(\boldsymbol{\mu}_i) \in A(\vartheta_i, \boldsymbol{\pi})$ . The Dirichlet formulation of the model is

$$\begin{aligned} Y_i &\stackrel{\text{ind}}{\sim} \sum_{j=1}^J \pi_j g(y_i \mid \mu_{ij}, \phi_{ij}), \\ \boldsymbol{\mu}_i &= \mathbf{V}^{(i)} \boldsymbol{\lambda}^{(i)}, \quad \text{where } \mathbf{V}^{(i)} \text{ contains vertices of } A(\vartheta_i, \boldsymbol{\pi}), \\ \boldsymbol{\lambda}^{(i)} &\stackrel{\text{ind}}{\sim} \text{Dirichlet}_{k_i}(\boldsymbol{\alpha}^{(i)}). \end{aligned} \tag{3.2}$$

We restrict  $\boldsymbol{\alpha}^{(i)}$  to the  $k_i$ -dimension vector  $\kappa \mathbf{1}$  so that all  $\boldsymbol{\lambda}^{(i)}$  follow a Symmetric Dirichlet distribution parameterized by a single scalar  $\kappa$ ; this is done for several reasons. First, the dimension  $k_i$  can vary with the observation so that an arbitrary  $\boldsymbol{\alpha}$  would not be compatible with all observations. Second, the ordering of the vertices in  $\mathbf{V}^{(i)}$  is somewhat arbitrary, and it is difficult to maintain a correspondence between individual vertices and the elements of  $\boldsymbol{\alpha}$ . Figure 2 plots the symmetric Dirichlet density for several  $\kappa$  when  $k = 3$ . Note that  $\kappa = 1$  corresponds to the uniform distribution on the simplex, while  $0 < \kappa < 1$  results in more density focused toward the vertices, and  $\kappa > 1$  focuses density toward the interior.

Now, to obtain a Mixture Link density based on the more practical Beta distribution, define  $\ell_{ij}$  and  $u_{ij}$  as the smallest and largest elements respectively of the  $j$ th row  $\mathbf{V}^{(i)}$ ; then  $(\ell_{ij}, u_{ij})$  forms the support of  $\mu_{ij}$ . The Beta formulation of the model is

$$\begin{aligned} Y_i &\stackrel{\text{ind}}{\sim} \sum_{j=1}^J \pi_j g(y_i \mid \mu_{ij}, \phi_{ij}), \\ \mu_{ij} &= (u_{ij} - \ell_{ij}) \psi_{ij} + \ell_{ij}, \quad j = 1, \dots, J, \\ \psi_{ij} &\sim \text{Beta}(a_{ij}, b_{ij}). \end{aligned} \tag{3.3}$$

To obtain  $a_{ij}$  and  $b_{ij}$ , we first compute

$$\mathbb{E}(\mu_{ij}) = (u_{ij} - \ell_{ij}) \frac{a_{ij}}{a_{ij} + b_{ij}} + \ell_{ij}, \quad \text{and} \quad \text{Var}(\mu_{ij}) = \frac{(u_{ij} - \ell_{ij})^2 a_{ij} b_{ij}}{(a_{ij} + b_{ij})^2 (a_{ij} + b_{ij} + 1)}.$$

Next, for  $\boldsymbol{\lambda} \sim \text{Dirichlet}_{k_i}(\kappa \mathbf{1})$  and  $\mathbf{v}_{j\cdot}^{(i)T}$  denoting the  $j$ th row of  $\mathbf{V}^{(i)}$ , we can obtain

$$\mathbb{E}(\mathbf{v}_{j\cdot}^{(i)T} \boldsymbol{\lambda}) = \bar{v}_{j\cdot}^{(i)} \quad \text{and} \quad \text{Var}(\mathbf{v}_{j\cdot}^{(i)T} \boldsymbol{\lambda}) = \frac{\mathbf{v}_{j\cdot}^{(i)T} \mathbf{v}_{j\cdot}^{(i)} - k_i (\bar{v}_{j\cdot}^{(i)})^2}{k_i (1 + k_i \kappa)},$$

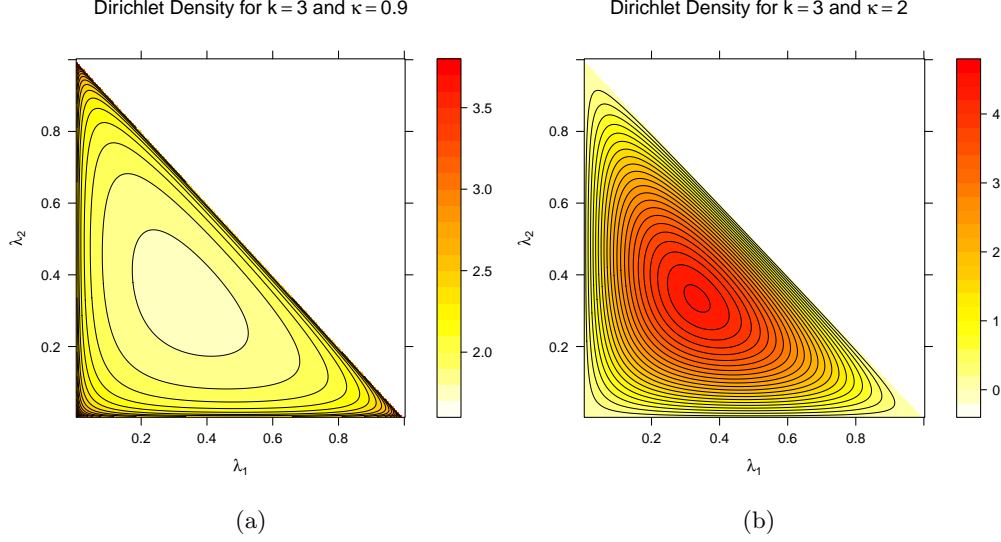


Figure 2: The  $\text{Dirichlet}_3(\boldsymbol{\lambda} \mid \kappa \mathbf{1})$  density for several settings of  $\kappa$ . Only  $\lambda_1$  and  $\lambda_2$  are plotted since  $\lambda_3 = 1 - \lambda_1 - \lambda_2$ .

where  $\bar{v}_{j\cdot}^{(i)}$  denotes the mean of  $\mathbf{v}_{j\cdot}^{(i)T}$ . Equating  $E(\mu_{ij})$  to  $E(\mathbf{v}_{j\cdot}^{(i)T} \boldsymbol{\lambda})$  and  $\text{Var}(\mu_{ij})$  to  $\text{Var}(\mathbf{v}_{j\cdot}^{(i)T} \boldsymbol{\lambda})$  and solving for  $a_{ij}$  and  $b_{ij}$ , we obtain that

$$a_{ij} = (\bar{v}_{j\cdot}^{(i)} - \ell_{ij})^2 \left[ \frac{k_i(1 + k_i\kappa)}{\mathbf{v}_{j\cdot}^{(i)T} \mathbf{v}_{j\cdot}^{(i)} - k_i(\bar{v}_{j\cdot}^{(i)})^2} \right] \frac{u_{ij} - \bar{v}_{j\cdot}^{(i)}}{u_{ij} - \ell_{ij}} - \frac{\bar{v}_{j\cdot}^{(i)} - \ell_{ij}}{u_{ij} - \ell_{ij}}, \quad (3.4)$$

$$b_{ij} = a_{ij} \left( \frac{u_{ij} - \bar{v}_{j\cdot}^{(i)}}{\bar{v}_{j\cdot}^{(i)} - \ell_{ij}} \right). \quad (3.5)$$

In the special case that  $k = 2$ , we have

$$\begin{aligned} \bar{v}_{j\cdot}^{(i)} &= \frac{1}{2} \left[ \min_{\ell \in \{1,2\}} v_{j\ell}^{(i)} + \max_{\ell \in \{1,2\}} v_{j\ell}^{(i)} \right] = \frac{1}{2} [\ell_{ij} + u_{ij}], \\ \bar{v}_{j\cdot}^{(i)} - \ell_{ij} &= u_{ij} - \bar{v}_{j\cdot}^{(i)}, \\ \mathbf{v}_{j\cdot}^{(i)T} \mathbf{v}_{j\cdot}^{(i)} &= u_{ij}^2 + \ell_{ij}^2, \end{aligned}$$

from which it can be shown that  $a_{ij} = \kappa$  and  $b_{ij} = \kappa$ .

Raim (2014) observes through simulation that, although the linear-combination-of-Dirichlet density can differ substantially from the moment-matched Beta density, the density of model (3.3) is a close approximation to the density of model (3.2). We have paid specific attention to the marginal distributions of the coordinates of  $\boldsymbol{\mu}_i$  rather than the full joint distribution; it is seen from (2.6) that only the marginals influence the overall Mixture Link distribution. The density of model (3.3) is now given by

$$f(y_i \mid \boldsymbol{\beta}, \boldsymbol{\pi}, \boldsymbol{\phi}_i, \kappa) = \sum_{j=1}^J \pi_j \int_0^1 g(y_i \mid H_{ij}(w), \boldsymbol{\phi}_{ij}) \cdot \mathcal{B}(w \mid a_{ij}, b_{ij}) dw, \quad (3.6)$$

where  $\mathcal{B}(x \mid a, b)$  denotes the Beta density and  $H_{ij}(x) = (u_{ij} - \ell_{ij})x + \ell_{ij}$ .

Computation of the Mixture Link density and its moments depends on the vertices of the set  $A$ . For the case  $J = 2$ , it is easy to identify the vertices of  $A$  graphically by plotting the line  $\mu_1\pi_1 + \mu_2\pi_2 = \vartheta$ , and

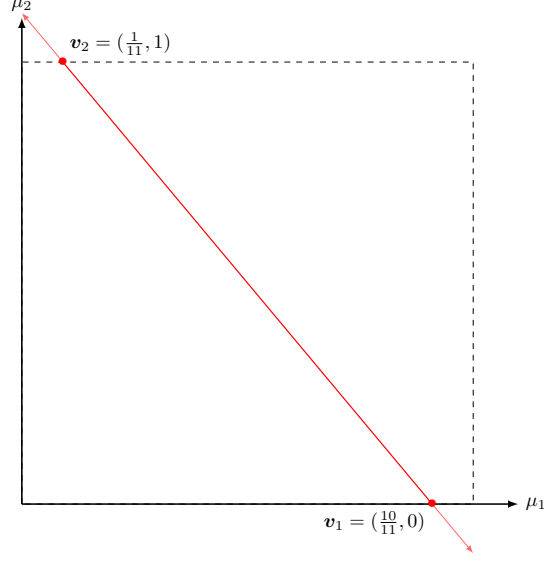


Figure 3: An illustration of the set  $A(\vartheta, \boldsymbol{\pi}) = \{\boldsymbol{\mu} \in [0, 1]^J : \boldsymbol{\mu}^T \boldsymbol{\pi} = \vartheta\}$ . Here we have selected  $\boldsymbol{\pi} = (\frac{11}{20}, \frac{9}{20})$  and  $\vartheta = \frac{1}{2}$ .

visually identifying the points at which it intersects the unit rectangle. An illustration is given in Figure 3. Formulas for the vertices in this case are stated now as a lemma.

**Lemma 3.1.** *Suppose  $J = 2$  and  $A$  has two distinct vertices  $\mathbf{v}_1, \mathbf{v}_2$ . Then the vertices are given by*

$$\mathbf{v}_1 = \begin{cases} \left(\frac{1}{\pi_1}\vartheta, 0\right), & \text{if } \frac{1}{\pi_1}\vartheta \leq 1 \\ \left(1, \frac{1}{\pi_2}(\vartheta - \pi_1)\right), & \text{otherwise,} \end{cases}$$

$$\mathbf{v}_2 = \begin{cases} \left(\frac{1}{\pi_1}(\vartheta - \pi_2), 1\right), & \text{if } \frac{1}{\pi_1}(\vartheta - \pi_2) \geq 0 \\ \left(0, \frac{1}{\pi_2}\vartheta\right), & \text{otherwise,} \end{cases}$$

where  $\pi_2 = 1 - \pi_1$ .

*Proof.* Using  $\mu_1\pi_1 + \mu_2\pi_2 = \vartheta$  we have

$$\mu_1 = \frac{1}{\pi_1}(\vartheta - \mu_2\pi_2) \quad \text{and} \quad \mu_2 = \frac{1}{\pi_2}(\vartheta - \mu_1\pi_1), \quad (3.7)$$

where  $\mu_1 \in [0, 1]$  and  $\mu_2 \in [0, 1]$  must hold. To obtain  $\mathbf{v}_1$ , take  $\mu_1$  as large as possible noting expressions (3.7). If  $\mu_1 = 1$  is a valid solution (i.e. a point in  $A$ ), then  $\mu_2 = \frac{1}{\pi_1}(\vartheta - \pi_2)$ . Otherwise, take  $\mu_2$  as small as possible to maximize  $\mu_1$ ; this yields  $\mu_1 = \frac{1}{\pi_1}\vartheta$  and  $\mu_2 = 0$ . A similar argument taking  $\mu_1$  as small as possible yields  $\mathbf{v}_2$ .  $\square$

We may also locate the vertices  $\mathbf{v}_1, \mathbf{v}_2$  systematically in the following way. Fix  $\mu_2 = 0$  and solve for  $\mu_1$  so that  $\boldsymbol{\mu}^T \boldsymbol{\pi} = \vartheta$ . Then fix  $\mu_2 = 1$  and solve for  $\mu_1$ . Then fix  $\mu_1$  at the values 0 and 1 and solve for  $\mu_2$ . At most two of these four solutions are contained in  $A$ ; these are the vertices. We will soon see that this idea generalizes to  $J > 2$ . Note that it is also possible to have  $k = 1$  vertices when  $J = 2$ . For example, if  $\boldsymbol{\pi} = (1/2, 1/2)$  and  $\vartheta = 1$ , then  $\mu_1 = 1, \mu_2 = 1$  is the only solution to  $\mu_1\pi_1 + \mu_2\pi_2 = \vartheta$  in  $[0, 1]^2$ , and therefore  $A$  is a singleton set.

For the general ( $J \geq 2$ ) case, Lemma 3.2 characterizes points in  $A$  which need to be considered when searching for the extreme points. In searching for extreme points, we must only consider those with at most one component not equal to 0 or 1.

**Lemma 3.2** (Characterization of Extreme Points of  $A$ ). *Suppose  $\mathbf{v} = (v_1, \dots, v_J)$  is a point in  $A$  with two or more components strictly between 0 and 1. Then  $\mathbf{v}$  is not an extreme point of  $A$ .*

*Proof.* Suppose without loss of generality that  $\mathbf{v} \in A$  with  $v_1 \in (0, 1)$  and  $v_2 \in (0, 1)$ . We have that

$$\begin{aligned} \mathbf{v}^T \boldsymbol{\pi} = \vartheta &\iff v_1 \pi_1 + v_2 \pi_2 + (v_3 \pi_3 + \dots + v_J \pi_J) = \vartheta \\ &\iff v_1 \pi_1 + v_2 \pi_2 = \vartheta^*, \end{aligned}$$

where  $\vartheta^* = \vartheta - (v_3 \pi_3 + \dots + v_J \pi_J)$ . We can now use Lemma 3.1 to obtain vertices, say  $\mathbf{a}$  and  $\mathbf{b}$ , of the line segment

$$L = \{(\mu_1, \mu_2, v_3, \dots, v_J) \in [0, 1]^J : \mu_1 \pi_1 + \mu_2 \pi_2 = \vartheta^*\},$$

where  $(v_3, \dots, v_J)$  are held fixed and only  $(\mu_1, \mu_2)$  may vary. Explicitly, we have

$$\begin{aligned} \mathbf{a} &= \begin{cases} \left(\frac{1}{\pi_1} \vartheta^*, 0, v_3, \dots, v_J\right), & \text{if } \frac{1}{\pi_1} \vartheta^* \leq 1 \\ \left(1, \frac{1}{\pi_2} (\vartheta^* - \pi_1), v_3, \dots, v_J\right), & \text{otherwise,} \end{cases} \\ \mathbf{b} &= \begin{cases} \left(\frac{1}{\pi_1} (\vartheta^* - \pi_2), 1, v_3, \dots, v_J\right), & \text{if } \frac{1}{\pi_1} (\vartheta^* - \pi_2) \geq 0 \\ \left(0, \frac{1}{\pi_2} \vartheta^*, v_3, \dots, v_J\right), & \text{otherwise.} \end{cases} \end{aligned}$$

By construction, we have that  $\mathbf{v}$  is in the line segment strictly between  $\mathbf{a}$  and  $\mathbf{b}$ , with  $\mathbf{a} \neq \mathbf{b}$ . Furthermore, since  $L \subseteq A$ , we have that  $\mathbf{a}, \mathbf{b} \in A$ . Therefore,  $\mathbf{v}$  can not be an extreme point of  $A$ .  $\square$

This can be used to formulate a simple procedure to identify all extreme points of  $A$ , which is given as Algorithm 3.1. Notice that it considers  $J \cdot 2^{J-1}$  points; this would be impractical for large  $J$ , but is manageable for smaller values of  $J$  that are commonly used in finite mixtures.

---

**Algorithm 3.1** Find vertices of the set  $A(\vartheta, \boldsymbol{\pi})$ .

---

```

function FINDVERTICES( $\vartheta, \boldsymbol{\pi}$ )
   $\mathcal{V} \leftarrow \emptyset$ 
  for  $j = 1, \dots, J$  do
    if  $\pi_j > 0$  then
      for all  $\boldsymbol{\mu}_{-j} \in \{0, 1\}^{J-1}$  do
         $\mu_j^* \leftarrow \pi_j^{-1} [\vartheta - \boldsymbol{\mu}_{-j}^T \boldsymbol{\pi}_{-j}]$ 
         $\mathbf{v}^* \leftarrow (\mu_1, \dots, \mu_{j-1}, \mu_j^*, \mu_{j+1}, \dots, \mu_J)$ 
         $\mathcal{V} \leftarrow \mathcal{V} \cup \mathbf{v}^*$  if  $\mathbf{v}^* \in A(\vartheta, \boldsymbol{\pi})$ 
  return Matrix  $\mathbf{V}$  with columns  $\mathbf{v}^* \in \mathcal{V}$ 

```

---

We will now formulate a Mixture Link Binomial distribution. Suppose  $g(y_i | w, \phi_{ij}) = \text{Bin}(y_i | m_i, w)$  so that  $y_i$  represents a count of successes out of  $m_i$  independent trials. Model (3.3) becomes

$$\begin{aligned} Y_i &\overset{\text{ind}}{\sim} \sum_{j=1}^J \pi_j \binom{m_i}{y_i} \mu_{ij}^{y_i} (1 - \mu_{ij})^{m_i - y_i}, \\ \mu_{ij} &= (u_{ij} - \ell_{ij}) \psi_{ij} + \ell_{ij}, \quad j = 1, \dots, J, \\ \psi_{ij} &\sim \text{Beta}(a_{ij}, b_{ij}). \end{aligned} \tag{3.8}$$

To draw from this distribution,

1. Compute matrix  $\mathbf{V}$  given  $\mathbf{x}$ ,  $\boldsymbol{\beta}$ , and  $\boldsymbol{\pi}$ .



2. Compute  $a_j$  and  $b_j$  for  $j = 1, \dots, J$  according to (3.5), and let  $(\ell_j, u_j)$  be the minimum and maximum element, respectively, of the  $j$ th row of  $\mathbf{V}$ .
3. Let  $\mu_j = (u_j - \ell_j)\psi_j + \ell_j$  with  $\psi_j \sim \text{Beta}(a_j, b_j)$ , for  $j = 1, \dots, J$ .
4. Draw  $Z \sim \text{Discrete}(1, \dots, J; \boldsymbol{\pi})$ .
5. Draw  $Y \sim \text{Binomial}(m, \mu_Z)$ .

Here,  $\text{Discrete}(1, \dots, k; \mathbf{p})$  denotes the discrete distribution with values  $1, \dots, k$  and corresponding probabilities  $\mathbf{p} = (p_1, \dots, p_k)$ . Moments of  $Y$  can be computed using moments of  $\mu_j$  for  $j = 1, \dots, J$ . In particular, after some algebra, we obtain

$$\text{Var}(Y) = m\vartheta(1 - m\vartheta) + m(m-1) \sum_{j=1}^J \pi_j \frac{\mathbf{v}_j^T \mathbf{v}_j + \kappa(k\bar{v}_j)^2}{k(1 + \kappa k)}.$$

Some remarks about the Mixture Link Binomial distribution follow.<sup>1</sup>

**Remark 3.3.** For the case  $m = 1$  where  $y$  represents a single success or failure,  $E(Y) = \vartheta$  implies  $P(Y = 1) = \vartheta^y(1 - \vartheta)^{1-y}$ , and Mixture Link simplifies to the usual Bernoulli regression model. In this case, the distribution depends only on its  $\boldsymbol{\beta}$  parameter. When  $m > 1$ , this trivial simplification does not take place.

**Remark 3.4.** Note that because  $\mathbf{v}_j^T \mathbf{v}_j \leq k$  and  $\bar{v}_j \leq 1$ , we have  $\sum_{j=1}^J \pi_j \mathbf{v}_j^T \mathbf{v}_j + \kappa(k\bar{v}_j)^2 \leq k(1 + \kappa k)$ , yielding the bound  $\text{Var}(Y) \leq m(m-1) - m\vartheta(m\vartheta - 1)$ , which is free of  $\boldsymbol{\pi}$  and  $\kappa$ .

**Remark 3.5.** The expression  $\text{Var}(Y)$  is non-increasing in  $\kappa$ . This can be seen from

$$\frac{\partial}{\partial \kappa} \text{Var}(Y) = -\frac{m(m-1)}{(1 + \kappa k)^2} \sum_{j=1}^J \pi_j \sum_{\ell=1}^k (v_{j\ell} - \bar{v}_j)^2 \leq 0.$$

**Remark 3.6.**  $\text{Binomial}(m, \vartheta)$  is a special case of Mixture Link Binomial, when  $\boldsymbol{\pi} = (\frac{1}{J}, \dots, \frac{1}{J})$  and  $\kappa \rightarrow \infty$ . This can be seen directly from the Dirichlet formulation of Mixture Link (3.2). Let  $\boldsymbol{\pi} = (\frac{1}{J}, \dots, \frac{1}{J})$  so that  $A(\boldsymbol{\pi}, \vartheta) = \{\boldsymbol{\mu} \in [0, 1]^J : \mu_1 + \dots + \mu_J = J\vartheta\}$ . A vertex  $\mathbf{v}^*$  of  $A(\boldsymbol{\pi}, \vartheta)$  is obtained by taking, say, the first  $v_1^*, \dots, v_{[J\vartheta]}^*$  to be 1,  $v_{[J\vartheta]+1}^* = J\vartheta - [J\vartheta]$ , and the remaining elements of  $\mathbf{v}^*$  to be zero. Here,  $[x]$  represents the integer part of a real number  $x$ . By Lemma 3.2,  $\mathbf{v}^*$  is a vertex of  $A(\boldsymbol{\pi}, \vartheta)$ . The remaining vertices can be obtained by permuting the elements of  $\mathbf{v}^*$ . If  $\tilde{v}_1^*, \dots, \tilde{v}_s^*$  are the unique elements of  $\mathbf{v}^*$  with multiplicities  $|\tilde{v}_1^*|, \dots, |\tilde{v}_s^*|$ , then there are  $k = J! / \{|\tilde{v}_1^*|! \dots |\tilde{v}_s^*|!\}$  unique permutations of  $\mathbf{v}^*$  to use as columns in the matrix  $\mathbf{V}$ . Notice that, for any  $a, j \in \{1, \dots, J\}$ , the element  $\tilde{v}_a^*$  appears in the  $j$ th row  $\mathbf{v}_j^T$  of  $\mathbf{V}$  exactly  $(J-1)! / \{|\tilde{v}_a^* - 1|! \prod_{\ell \neq a} |\tilde{v}_\ell^*|!\}$  times.<sup>2</sup> Then we have

$$\mathbf{v}_j^T \mathbf{1} = \sum_{a=1}^s \tilde{v}_a^* \frac{(J-1)!}{|\tilde{v}_a^* - 1|! \prod_{\ell \neq a} |\tilde{v}_\ell^*|!} = \sum_{a=1}^s \tilde{v}_a^* \frac{J! |\tilde{v}_a^*|}{\prod_{\ell=1}^a |\tilde{v}_\ell^*|!} \frac{1}{J} = \frac{k}{J} \sum_{a=1}^s \tilde{v}_a^* \cdot |\tilde{v}_a^*| = \frac{k}{J} J\vartheta = k\vartheta. \quad (3.9)$$

When  $\kappa \rightarrow \infty$ , a draw  $\boldsymbol{\lambda} \sim \text{Dirichlet}_k(\kappa \mathbf{1})$  becomes a point mass at its expected value  $\frac{1}{k} \mathbf{1}$  so that (3.9) gives  $\boldsymbol{\mu} = \mathbf{V}\boldsymbol{\lambda} = \frac{1}{k} \mathbf{V}\mathbf{1} = \vartheta \mathbf{1}$ . It can now be seen that

$$f(y) = \sum_{j=1}^J \pi_j \binom{m}{y} \mu_j^y (1 - \mu_j)^{m-y} = \sum_{j=1}^J \frac{1}{J} \binom{m}{y} \vartheta^y (1 - \vartheta)^{m-y}$$

is the  $\text{Binomial}(m, \vartheta)$  distribution.

<sup>1</sup>Analogous statements for some of these remarks can be made about the Mixture Link Poisson and Mixture Link Normal distributions, discussed in Sections 4 and 5. We have focused on the Binomial case for brevity.

<sup>2</sup>This is the number of unique permutations of  $\{v_1^*, \dots, v_J^*\}$ , keeping one of the elements fixed.

**Remark 3.7.** Mixture Link Binomial becomes a zero- and/or  $m$ -inflated Binomial model when  $\kappa \rightarrow 0$ . As in Remark 3.6, we will work directly from the Dirichlet formulation. As  $\kappa \rightarrow 0$ , a draw  $\boldsymbol{\lambda} \sim \text{Dirichlet}_k(\kappa \mathbf{1})$  behaves as a discrete uniform random variable on  $\{\mathbf{e}_1, \dots, \mathbf{e}_k\}$ , the columns of the  $k \times k$  identity matrix which represent the vertices of the simplex  $\mathcal{S}^k$ . Here, the Mixture Link distribution becomes

$$\begin{aligned} f(y) &= \sum_{j=1}^J \pi_j \sum_{\ell=1}^k \frac{1}{k} \cdot \text{Bin}(y \mid m, \mathbf{v}_j^T \mathbf{e}_\ell) \\ &= \sum_{j=1}^J \sum_{\ell=1}^k \frac{\pi_j}{k} \binom{m}{y} v_{j\ell}^y (1 - v_{j\ell})^{m-y}. \end{aligned}$$

Recall from Lemma 3.2 that, for each  $\ell = 1, \dots, k$ , at most one of  $\{v_{1\ell}, \dots, v_{J\ell}\}$  can take on a value outside of  $\{0, 1\}$ . Terms with  $v_{j\ell} = 0$  represent a point mass at zero, while terms with  $v_{j\ell} = 1$  represent a point mass at  $m$ .

**Remark 3.8.** Mixture Link Binomial is closely related to two other Binomial models for overdispersion. Starting from (3.6), if we could take  $\ell_{ij} = 0$  and  $u_{ij} = 1$ , we would have

$$\begin{aligned} f(y_i \mid \boldsymbol{\beta}, \boldsymbol{\pi}, \boldsymbol{\phi}_i, \kappa) &= \sum_{j=1}^J \pi_j \int_0^1 \text{Bin}(y_i \mid (u_{ij} - \ell_{ij})w + \ell_{ij}, \boldsymbol{\phi}_{ij}) \cdot \mathcal{B}(w \mid a_{ij}, b_{ij}) dw, \\ &= \sum_{j=1}^J \pi_j \binom{m_i}{y_i} \frac{B(a_{ij} + y_i, b_{ij} + m_i - y_i)}{B(a_{ij}, b_{ij})}. \end{aligned}$$

Therefore, Mixture Link Binomial can be seen as a constrained form of a finite mixture of  $J$  Beta-Binomial densities. Also, recall the Random-Clumped Binomial (RCB) distribution (Morel and Nagaraj, 1993), whose density is given by

$$f(y \mid \pi, \rho) = \pi_1 \text{Bin}(y \mid \pi, \mu_1) + \pi_2 \text{Bin}(y \mid \pi, \mu_2),$$

where  $\pi_1 = \pi$ ,  $\pi_2 = 1 - \pi$ , and  $\mu_1 = (1 - \rho)\pi + \rho$ ,  $\mu_2 = (1 - \rho)\pi$ . The free parameters of the distribution are  $\pi \in (0, 1)$  and  $\rho \in (0, 1)$ . Notice that  $\pi_1 \mu_1 + \pi_2 \mu_2 = \pi$ , so that this particular choice of  $(\mu_1, \mu_2)$  is in the set  $A(\pi_1, \boldsymbol{\pi})$ . Therefore, RCB can be seen as a special case of Mixture Link Binomial.

## 4 Positive Means

The setting  $\mathcal{M} = [0, \infty)$  is commonly required for count data and time-to-event data. Just as in Section 3, the set  $A(\vartheta, \boldsymbol{\pi}) = \{\boldsymbol{\mu} \in [0, \infty)^J : \boldsymbol{\mu}^T \boldsymbol{\pi} = \vartheta\}$  is a closed convex hyperplane segment within  $\mathbb{R}^J$ . Therefore, the decomposition (3.1) also applies but the procedure to compute vertices is much simpler. First note that for  $J = 2$ ,  $\mathbf{v}_1 = (\vartheta/\pi_1, 0)$  and  $\mathbf{v}_2 = (0, \vartheta/\pi_2)$  are the vertices of  $A$ . To see this, suppose  $\boldsymbol{\mu}^*$  is an arbitrary point in  $A$ . Then we must have, for some  $\lambda \in [0, 1]$ ,

$$\begin{pmatrix} \mu_1^* \\ \mu_2^* \end{pmatrix} = \lambda \mathbf{v}_1 + (1 - \lambda) \mathbf{v}_2 = \begin{pmatrix} \lambda \vartheta / \pi_1 \\ (1 - \lambda) \vartheta / \pi_2 \end{pmatrix}.$$

Taking  $\lambda = \mu_1^* \pi_1 / \vartheta$  satisfies the first equation  $\mu_1^* = \lambda \vartheta / \pi_1$ , and also gives  $(1 - \lambda) \vartheta / \pi_2 = (\vartheta - \mu_1^* \pi_1) / \pi_2 = \mu_2^*$  to satisfy the second equation. Similarly to Lemma 3.2, we characterize the extreme points of  $A$  for the case of positive means by Lemma 4.1. The proof is similar to that of Lemma 3.2, and therefore omitted.

**Lemma 4.1** (Characterization of Extreme Points of  $A$ ). *Suppose  $\mathbf{v} = (v_1, \dots, v_J)$  is a point in  $A$  with two or more components which are strictly positive. Then  $\mathbf{v}$  is not an extreme point of  $A$ .*

Now, if  $\mathbf{v} = (0, \dots, 0, v_j, 0, \dots, 0)$  is a point in  $A$ ,  $\mathbf{v}^T \boldsymbol{\pi} = \vartheta$  implies  $v_j \pi_j = \vartheta$ . There are exactly  $J$  such points in  $A$ , yielding  $\mathbf{V} = \text{Diag}(\vartheta/\pi_1, \dots, \vartheta/\pi_J)$ . Poisson Mixture Link can now be formulated similarly as in Section 3. Note that, in this case, the Dirichlet and Beta assumptions on  $\mu_i$  lead to exactly the same model. Taking  $g(y_i | w, \phi_{ij}) = \text{Poisson}(y_i | w)$ , the model becomes

$$\begin{aligned} Y_i &\stackrel{\text{ind}}{\sim} \sum_{j=1}^J \pi_j \frac{e^{-\mu_{ij}} \mu_{ij}^{y_i}}{y_i!} \\ \boldsymbol{\mu}_i &= \mathbf{V}^{(i)} \boldsymbol{\lambda}^{(i)}, \\ \boldsymbol{\lambda}^{(i)} &\stackrel{\text{ind}}{\sim} \text{Dirichlet}_{k_i}(\boldsymbol{\kappa} \mathbf{1}). \end{aligned}$$

Expressions involving the vertices simplify in the case of positive means, with  $J = k_i$ ,  $\ell_{ij} = 0$ ,  $u_{ij} = v_{jj}^{(i)}$ ,  $\bar{v}_j^{(i)} = v_{jj}^{(i)}/J$ ,  $\mathbf{v}_j^{(i)T} \mathbf{v}_j^{(i)} = (v_{jj}^{(i)})^2$ ,  $H_{ij}(w) = v_{jj}^{(i)} w$ ,  $a_{ij} = \kappa$ , and  $b_{ij} = \kappa(J-1)$ . Recalling that the marginal distribution of a single coordinate of  $\text{Dirichlet}_J(\boldsymbol{\kappa} \mathbf{1})$  is  $\text{Beta}(\kappa, \kappa(J-1))$ , the Mixture Link density becomes

$$\begin{aligned} f(y_i | \boldsymbol{\beta}, \boldsymbol{\pi}, \kappa) &= \sum_{j=1}^J \pi_j \int_0^1 \frac{e^{-H_{ij}(w)} H_{ij}(w)^{y_i}}{y_i!} \cdot \mathcal{B}(w | \kappa, \kappa(J-1)) dw \\ &= \sum_{j=1}^J \pi_j \int_0^1 \frac{e^{-v_{jj}^{(i)} w} [v_{jj}^{(i)} w]^{y_i}}{y_i!} \cdot \frac{w^{\kappa-1} (1-w)^{\kappa(J-1)-1}}{B(\kappa, \kappa(J-1))} dw \\ &= \frac{\vartheta^{y_i} \Gamma(y_i + \kappa) \Gamma(\kappa J)}{\Gamma(y_i + \kappa J) \Gamma(\kappa) \Gamma(y_i + 1)} \sum_{j=1}^J \pi_j^{1-y_i} \cdot \mathcal{F}\left(-\frac{\vartheta_i}{\pi_j}; y_i + \kappa, y_i + J\kappa\right) \end{aligned}$$

where  $\mathcal{F}(x; a, b) = [B(a, b-a)]^{-1} \int_0^1 w^{a-1} (1-w)^{b-a-1} e^{xw} dw$  is the confluent hypergeometric function of the first order and  $B(a, b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$  is the beta function (Johnson et al., 2005, Chapter 1). Implementations of  $\mathcal{F}(x; a, b)$  are available in computing packages such as the GNU Scientific Library.<sup>3</sup> The variance of  $Y$  becomes

$$\begin{aligned} \text{Var}(Y) &= \vartheta + \left[ \sum_{j=1}^J \pi_j \bar{v}_j^2 - \vartheta^2 \right] + \sum_{j=1}^J \pi_j \frac{\mathbf{v}_j^T \mathbf{v}_j - k(\bar{v}_j)^2}{k(1 + \kappa k)} \\ &= \vartheta + \vartheta^2 \left[ \frac{\kappa + 1}{J(1 + J\kappa)} \sum_{j=1}^J \frac{1}{\pi_j} - 1 \right]. \end{aligned}$$

Drawing random variables from Mixture Link Poisson is similar to the method given in Section 3 for Mixture Link Binomial:

1. Compute matrix of vertices  $\mathbf{V}$  given  $\mathbf{x}$ ,  $\boldsymbol{\beta}$ , and  $\boldsymbol{\pi}$ .
2. Let  $\mu_j = \psi_j \cdot \vartheta/\pi_j$  with  $\psi_j \sim \text{Beta}(\kappa, \kappa(J-1))$ , for  $j = 1, \dots, J$ .
3. Draw  $Z \sim \text{Discrete}(1, \dots, J; \boldsymbol{\pi})$ .
4. Draw  $Y \sim \text{Binomial}(m, \mu_Z)$ .

**Remark 4.2.** The expression  $\text{Var}(Y)$  is decreasing in  $\kappa$  since

$$\frac{\partial}{\partial \kappa} \text{Var}(Y) = -\frac{\vartheta(J-1)}{J(1+J\kappa)} \sum_{j=1}^J \frac{1}{\pi_j} < 0.$$

---

<sup>3</sup>[www.gnu.org/software/gsl](http://www.gnu.org/software/gsl)

## 5 Real-valued Means

In the case  $\mathcal{M} = \mathbb{R}$ , the set  $A(\vartheta, \boldsymbol{\pi}) = \{\boldsymbol{\mu} \in \mathbb{R}^J : \boldsymbol{\mu}^T \boldsymbol{\pi} = \vartheta\}$  forms a hyperplane in  $\mathbb{R}^J$  and can be decomposed as  $A(\vartheta, \boldsymbol{\pi}) = \{\bar{\boldsymbol{\mu}} \in \mathbb{R}^J : \bar{\boldsymbol{\mu}}^T \boldsymbol{\pi} = 0\} + \vartheta \mathbf{1}$ . For any  $\bar{\boldsymbol{\mu}}$  in the subspace  $\{\bar{\boldsymbol{\mu}} \in \mathbb{R}^J : \bar{\boldsymbol{\mu}}^T \boldsymbol{\pi} = 0\}$ , we can write  $\bar{\boldsymbol{\mu}}_J = -\pi_J^{-1}(\pi_1 \bar{\mu}_1 + \dots + \pi_{J-1} \bar{\mu}_{J-1})$  with  $\bar{\mu}_j$  unrestricted for  $j = 1, \dots, J-1$ . Therefore a basis for the subspace is given by the  $J \times (J-1)$  matrix

$$\mathbf{V} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ & & \ddots & \\ 0 & 0 & \cdots & 1 \\ -\pi_1/\pi_J & -\pi_2/\pi_J & \cdots & -\pi_{J-1}/\pi_J \end{pmatrix}.$$

We can therefore represent any  $\boldsymbol{\mu} \in A(\vartheta, \boldsymbol{\pi})$  as

$$\boldsymbol{\mu} = \mathbf{V}\boldsymbol{\lambda} + \vartheta \mathbf{1} \quad \text{for some } \boldsymbol{\lambda} \in \mathbb{R}^{J-1}.$$

A natural choice for a random effects distribution on  $A(\vartheta, \boldsymbol{\pi})$  is to take  $\lambda_j \stackrel{\text{iid}}{\sim} \text{N}(0, \kappa^2)$  for  $j = 1, \dots, J-1$ . This leads to

$$\boldsymbol{\mu} \sim \text{N}(\vartheta \mathbf{1}, \kappa^2 \mathbf{V}\mathbf{V}^T), \quad \text{where} \quad \mathbf{V}\mathbf{V}^T = \begin{pmatrix} \mathbf{I} & -\pi_J^{-1} \boldsymbol{\pi}_{-J} \\ -\pi_J^{-1} \boldsymbol{\pi}_{-J}^T & \pi_J^{-2} \boldsymbol{\pi}_{-J}^T \boldsymbol{\pi}_{-J} \end{pmatrix},$$

$\mathbf{I}$  denotes the  $(J-1) \times (J-1)$  identity matrix, and  $\boldsymbol{\pi}_{-J} = (\pi_1, \dots, \pi_{J-1})$ . The Mixture Link density depends only on the diagonal terms of the random effect variance,

$$f(y_i | \boldsymbol{\beta}, \boldsymbol{\pi}, \phi_i, \kappa) = \sum_{j=1}^J \pi_j \int g(y_i | w, \phi_{ij}) \cdot \text{N}(w | \vartheta_i, \kappa^2 a_{ij}) dw, \quad (5.1)$$

where  $a_{ij} = 1$  for  $j = 1, \dots, J-1$  and  $a_{iJ} = \pi_J^{-2} \boldsymbol{\pi}_{-J}^T \boldsymbol{\pi}_{-J}$ .

To obtain a Mixture Link analogue to the commonly used ordinary least squares model, suppose  $g(y_i | w, \phi_{ij}) = \text{N}(y_i | w, \sigma_j^2)$ . In this case, it can be shown that (5.1) simplifies to the finite mixture

$$f(y_i | \boldsymbol{\beta}, \boldsymbol{\pi}, \sigma_1^2, \dots, \sigma_J^2, \kappa) = \sum_{j=1}^J \pi_j \text{N}(y_i | \vartheta_i, \kappa^2 a_{ij} + \sigma_j^2), \quad (5.2)$$

where each of the subpopulations has a common mean. If the  $J$  subpopulations are assumed to be homoskedastic, (5.2) further simplifies to a finite mixture of two densities,

$$f(y_i | \boldsymbol{\beta}, \boldsymbol{\pi}, \sigma^2, \kappa) = (1 - \pi_J) \text{N}(y_i | \vartheta_i, \kappa^2 + \sigma^2) + \pi_J \text{N}(y_i | \vartheta_i, \kappa^2 \pi_J^{-2} (1 - \pi_J)^2 + \sigma^2).$$

Focusing on the homoskedastic model, it is straightforward to draw from the distribution:

1. Draw  $Z_i \sim \text{Discrete}(1, 2; (1 - \pi_J, \pi_J))$ ,
2. Draw  $Y_i$  from  $\text{N}(y_i | \vartheta_i, \kappa^2 a_{ij} + \sigma^2)$  where  $Z_i = j$ .

An expression for the variance is given by

$$\text{Var}(Y_i) = \kappa^2 \frac{1 - \pi_J}{\pi_J} + \sigma^2.$$

## 6 Data Analysis Examples

We now present two examples of data analysis with the Mixture Link distribution. The Hiroshima data discussed in Section 6.1 features a Binomial outcome. The Arizona Medpar data has a count outcome, and is discussed in Section 6.2.

For a complete Bayesian specification of Mixture Link Binomial and Mixture Link Poisson, we assume priors

$$\begin{aligned}\boldsymbol{\beta} &\sim \text{N}(\mathbf{0}, \boldsymbol{\Omega}_{\boldsymbol{\beta}}), \\ \boldsymbol{\pi} &\sim \text{Dirichlet}(\boldsymbol{\gamma}), \\ \kappa &\sim \text{Gamma}(a_{\kappa}, b_{\kappa}),\end{aligned}$$

where the parameterization of Gamma is taken to have  $E(\kappa) = a_{\kappa}/b_{\kappa}$ . In the absence of a-priori knowledge, a somewhat vague choice of hyperparameters is  $\boldsymbol{\Omega}_{\boldsymbol{\beta}} = 1000\mathbf{I}_d$ ,  $\boldsymbol{\gamma} = \mathbf{1}$ , and  $a_{\kappa} = 1, b_{\kappa} = 2$ .

To diagnose the fit of models with non-Normal outcomes, we make use of the randomized quantile residuals (Dunn and Smyth, 1996). Interpretation of quantile residuals is similar to the routine residual analysis from ordinary least squares regression. Quantile residuals from an adequate model fit appear to behave as an independent sample from the standard Normal distribution. For  $y_i$  drawn independently from a continuous distribution  $F(\cdot | \boldsymbol{\theta})$  with estimate  $\hat{\boldsymbol{\theta}}$ , the quantile residual is defined as  $r_i = \Phi^{-1}\{F(y_i | \hat{\boldsymbol{\theta}})\}$ . For  $y_i$  drawn independently from a discrete distribution, there is an additional randomization where the residual is defined by  $r_i = \Phi^{-1}\{u_i\}$ , using  $u_i$  drawn uniformly on the interval between  $\lim_{\varepsilon \downarrow 0} F(y_i - \varepsilon | \hat{\boldsymbol{\theta}})$  and  $F(y_i | \hat{\boldsymbol{\theta}})$ . A Bayesian version of the quantile residual using draws  $\boldsymbol{\theta}^{(1)}, \dots, \boldsymbol{\theta}^{(R)}$  from the posterior distribution  $f(\boldsymbol{\theta} | \mathbf{y})$  is  $r_i = \frac{1}{R} \sum_{r=1}^R \Phi^{-1}\{u_i^{(r)}\}$ , where each  $u_i^{(r)}$  is drawn uniformly on the interval between  $\lim_{\varepsilon \downarrow 0} F(y_i - \varepsilon | \boldsymbol{\theta}^{(r)})$  and  $F(y_i | \boldsymbol{\theta}^{(r)})$ .

We will also evaluate models using prediction intervals computed from the posterior predictive distribution. Recall that the posterior predictive distribution for a new sample  $\tilde{\mathbf{y}}$  given the observed sample  $\mathbf{y}$  is

$$f(\tilde{\mathbf{y}} | \mathbf{y}) = \int f(\tilde{\mathbf{y}} | \boldsymbol{\theta}, \mathbf{y}) f(\boldsymbol{\theta} | \mathbf{y}) d\nu(\boldsymbol{\theta}) = \int f(\tilde{\mathbf{y}} | \boldsymbol{\theta}) f(\boldsymbol{\theta} | \mathbf{y}) d\nu(\boldsymbol{\theta}),$$

where  $\nu$  denotes an appropriate dominating measure. Then to sample from  $f(\tilde{\mathbf{y}} | \mathbf{y})$ :

1. Draw  $\boldsymbol{\theta}^{(1)}, \dots, \boldsymbol{\theta}^{(R)}$  from posterior  $f(\boldsymbol{\theta} | \mathbf{y})$ .
2. Draw  $\tilde{\mathbf{y}}^{(r)}$  from  $f(\tilde{\mathbf{y}} | \boldsymbol{\theta}^{(r)})$  for  $r = 1, \dots, R$ .

Now  $(\tilde{\mathbf{y}}^{(1)}, \dots, \tilde{\mathbf{y}}^{(R)})$  is a draw from the posterior predictive distribution. A prediction for the  $i$ th observation is given by  $\frac{1}{R} \sum_{r=1}^R \tilde{y}_i^{(r)}$ , and a prediction interval with coverage probability  $1 - \alpha$  for the  $i$ th observation is given by the  $\alpha/2$  and  $1 - \alpha/2$  quantiles of  $(\tilde{y}_i^{(1)}, \dots, \tilde{y}_i^{(R)})$ .

Label switching is a common issue in Bayesian analysis of finite mixtures (Jasra et al., 2005). For Mixture Link, the  $\boldsymbol{\pi}$  parameters are susceptible to this problem. Because finite mixtures are invariant to permutation of the labels, the parameters corresponding to labels  $\{1, \dots, J\}$  can change during the course of an MCMC computation. Therefore, special care must be taken when summarizing parameters using MCMC draws. In this work, we take the simple approach of reordering the components within each draw  $\boldsymbol{\pi}^{(r)}$ , in ascending order, for each  $r = 1, \dots, R$ .

### 6.1 Hiroshima Data

Awa et al. (1971) and Sofuni et al. (1978) study the effects of radiation exposure on chromosome aberrations in survivors of the atomic bombs that were used in Hiroshima and Nagasaki. We consider a subset of the data, as presented in Morel and Neerchal (2012), on  $n = 648$  subjects in Hiroshima. For the  $i$ th subject, a chromosome analysis has been carried out on  $m_i$  circulating lymphocytes to determine the number  $y_i$

Table 1: DIC for Hiroshima models.

Model	DIC
Binomial	3625.34
RCB	3148.05
BB	2984.49
MixLinkJ2	2876.64
MixLinkJ3	2878.01
MixLinkJ4	2875.93

containing chromosome aberrations. Neutron and gamma radiation exposure (measured in rads) are available as potential covariates. As in [Raim et al. \(2015\)](#), we consider the regression

$$\vartheta_i = G(\beta_0 + \beta_1 x_i + \beta_2 x_i^2), \quad (6.1)$$

where  $x_i$  is a normalized sum of neutron and gamma doses, and we take  $G$  to be the logistic CDF (as in logistic regression).

We compare six Binomial-type models with (6.1) as the regression function: Binomial, Random-Clumped Binomial (RCB), Beta-Binomial (BB), and Mixture Link with  $J = 2, 3, 4$  mixture components (MixLinkJ2, MixLinkJ3, MixLinkJ4). Because of the complicated manner in which parameters enter the Mixture Link Binomial likelihood, conjugate priors leading to closed-form Gibbs samplers do not appear possible. We considered a simple Random Walk Metropolis-Hastings (RWMH) sampler ([Robert and Casella, 2010](#), Section 7.5); however, sampling with RWMH is time consuming because it requires computation of the likelihood to determine whether each proposed jump will be accepted. Recall that, for Mixture Link Binomial, evaluation of the likelihood consists of evaluating  $J$  integrals numerically for each of the  $n$  observations. Alternatively, Appendix A proposes a Metropolis-within-Gibbs (MWG) sampler ([Robert and Casella, 2010](#), Section 10.3) where  $\psi_i$  are taken as augmented data ([Tanner and Wong, 1987](#)) to avoid the expensive integration.

An RWMH sampler was used to obtain posterior draws under the Binomial, RCB, and BB models, while the MWG sampler from Appendix A was used for Mixture Link. For each Mixture Link model, we carried out a preliminary “pilot” MCMC, which was used to tune the proposal distribution for a final MCMC run and achieve satisfactory mixing. Mixing was assessed primarily through trace plots and autocorrelation plots of the saved draws. Trace plots for the selected Mixture Link model are shown in Figure 6. For all models, a multivariate Normal proposal distribution was selected by hand to achieve acceptance rates between about 15% and 30%. Final MCMC runs for Mixture Link were carried out for 55,000 iterations; the first 5,000 were discarded as a burn-in sample, and 1 of every 50 remaining draws from the chain were saved. For Binomial, BB, and RCB, we used 50,000 iterations overall with the first 5,000 discarded as burn-in and saved 1 of every 50 remaining.

Table 1 shows the Deviance Information Criterion (DIC) for these models. The three Mixture Link models fit best according to DIC; BB has a smaller DIC than RCB by a large margin, and Binomial gives the worst fit as expected. Table 2 reports means, standard deviations, 2.5% quantiles, and 97.5% quantiles for each parameter from the posterior draws. Generally, signs and magnitudes of the  $\beta$  estimates agree between models. Standard deviations and credible intervals are a bit larger for BB and MixLink models than RCB and Binomial. Figure 4 displays quantile residuals for the Binomial, BB, and MixLinkJ2 models. Residuals from BB and MixLinkJ2 are markedly closer to a  $N(0, 1)$  sample than Binomial residuals, as can be seen from the Q-Q plots. For all models, there is a systematic pattern in residuals vs. predicted proportions, which is an indication that the mean is not fully explained by regression function (6.1). Finally, Figure 5 plots  $x_i$  against observed  $y_i/m_i$ , along with 95% prediction intervals for Binomial, BB, and MixLinkJ2. The intervals computed by MixLinkJ2, and to a lesser extent BB, express variability from the observed data into wider prediction intervals.

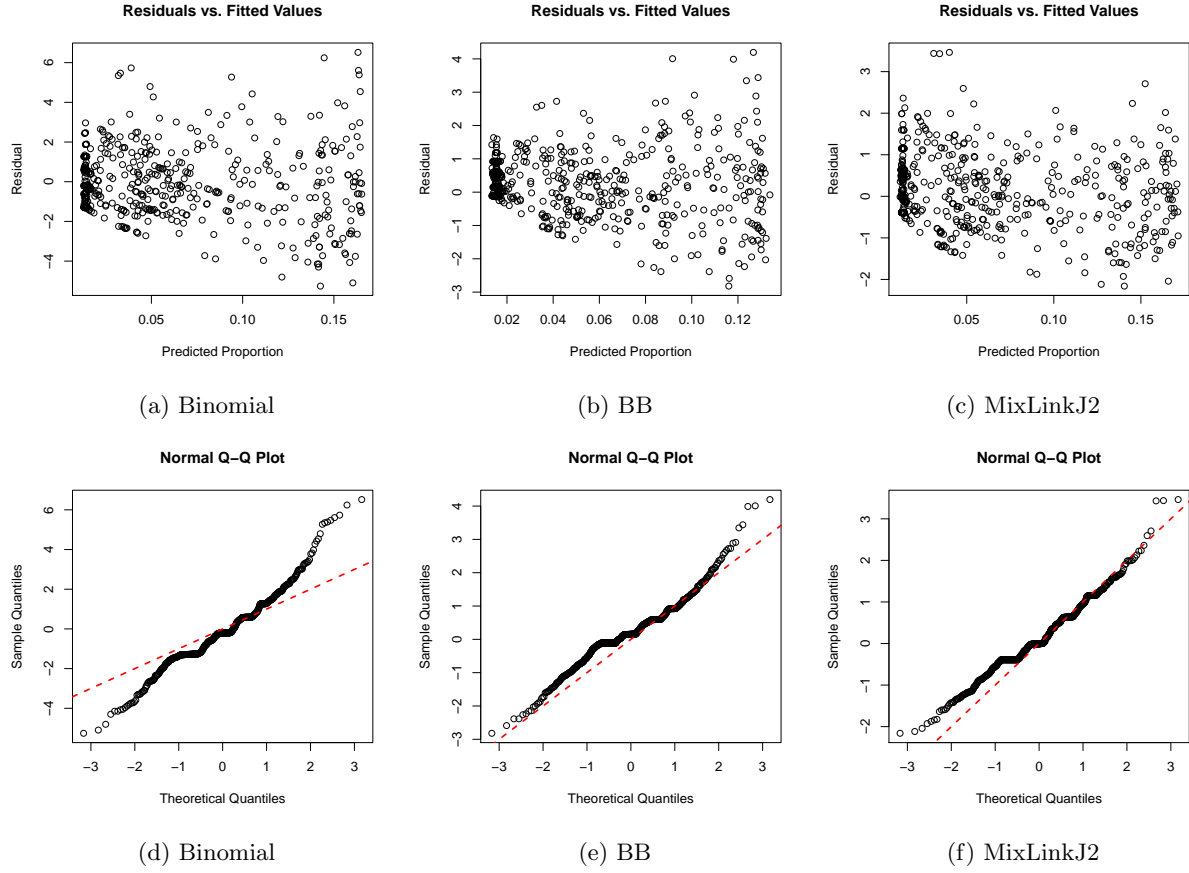


Figure 4: Quantile residuals for Hiroshima models.

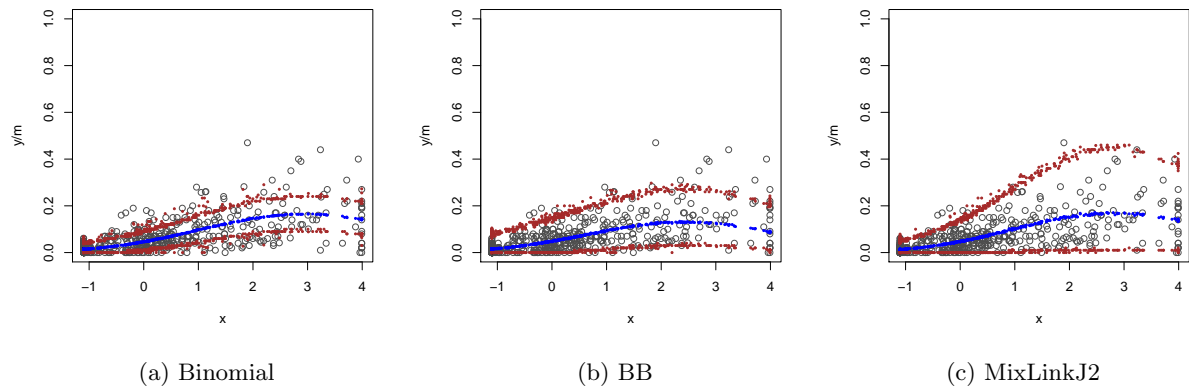


Figure 5: Observed proportions  $y_i/m_i$  vs.  $x_i$  for Hiroshima data are plotted as open circles. Smaller solid dots represent 95% prediction intervals (upper and lower curves) and predictions (middle curve) from the respective model.

Table 2: Posterior summaries for Hiroshima models.

Binomial	mean	SD	2.5%	97.5%
intercept	-3.0241	0.0241	-3.0695	-2.9723
$x$	0.9494	0.0244	0.9014	0.9938
$x^2$	-0.1611	0.0080	-0.1762	-0.1459
BB	mean	SD	2.5%	97.5%
intercept	-2.9437	0.0461	-3.0368	-2.8589
$x$	0.8165	0.0395	0.7346	0.8950
$x^2$	-0.1416	0.0139	-0.1681	-0.1146
$\rho$	0.1666	0.0079	0.1515	0.1823
RCB	mean	SD	2.5%	97.5%
intercept	-2.9761	0.0360	-3.0449	-2.9051
$x$	0.8859	0.0298	0.8296	0.9430
$x^2$	-0.1817	0.0121	-0.2052	-0.1578
$\rho$	0.1526	0.0081	0.1366	0.1678
MixLinkJ2	mean	SD	2.5%	97.5%
intercept	-3.0030	0.0440	-3.0857	-2.9110
$x$	0.9989	0.0426	0.9155	1.0880
$x^2$	-0.1771	0.0167	-0.2114	-0.1450
$\pi_1$	0.3336	0.0178	0.3004	0.3687
$\pi_2$	0.6664	0.0178	0.6313	0.6996
$\kappa$	1.6200	0.2489	1.2154	2.1959

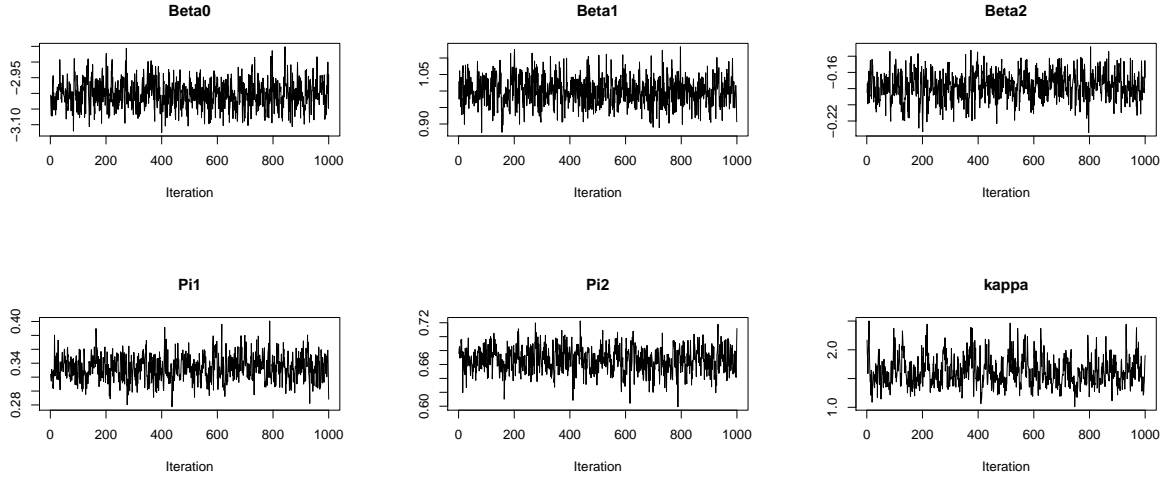


Figure 6: Trace plots for MixLinkJ2 fit to Hiroshima data.



## 6.2 Arizona Medpar Data

The `azpro` data in the `COUNT` R package are taken from Arizona cardiovascular patient files in 1991. It contains 3,589 observations on subjects from 17 hospitals. The outcome of interest, length of hospital stay  $y$ , is a count. Several indicator variables are available as covariates: `procedure` takes values 1 for Coronary Artery Bypass Graft and 0 for Percutaneous Transluminal Coronary Angioplasty, `sex` is 1 for male and 0 for female, type of admission `admit` is 1 if emergency and 0 if elective, `age75` is 1 if patient's age is at least 75 and 0 otherwise, and `hospital` is a code to identify hospital. For this example, we consider only the 376 observations with `hospital` = 6.5, and take the regression function to be

$$E(y_i) = \exp\{\beta_0 + \beta_1 \cdot \text{procedure}_i + \beta_2 \cdot \text{sex}_i + \beta_3 \cdot \text{age75}_i\}.$$

We compare count regression models based on Poisson, NegBin, and Mixture Link with  $J = 2, \dots, 8$  mixture components. All models used a simple RWMH sampler to obtain draws from the posterior. For Mixture Link models, proposals for  $\theta$  were drawn in a partitioned manner to improve mixing of the chain: a proposal for either  $\beta$ ,  $\pi$ , or  $\kappa$  was drawn at a time, keeping other parameters fixed, and either accepted or rejected. In some cases where  $J > 2$ , the components of  $\pi$  were also drawn individually to further improve mixing. We assessed mixing primarily through trace plots and autocorrelation plots of the saved draws. For all models, the multivariate Normal proposal distribution was tuned by hand to achieve acceptance rates between about 15% and 30%. MCMC was carried out for 55,000 iterations; the first 5,000 were discarded as a burn-in sample, and 1 of every 20 remaining draws from the chain were saved.

Table 3 compares DIC across all fitted models. Because Poisson is a special case of NegBin, it is not surprising that the DIC of NegBin indicates a superior fit. It is interesting that the DIC of MixLink appears to improve gradually as the number of mixture components  $J$  are increased. Taking  $J > 2$  required additional hand-tuning of the sampler for some cases to yield acceptable diagnostics. Initial attempts to fit MixLink with  $J = 9$  resulted in poor diagnostics, so these results are not shown. Figure 9 displays the trace plots for MixLinkJ8, which was selected among the seven Mixture Link models for further analysis.

We proceed by comparing the Poisson, NegBin, and MixLinkJ8 models. Table 4 reports means, standard deviations, 2.5% quantiles, and 97.5% quantiles of each parameter computed from the posterior draws. Generally, the signs and magnitudes of the means of  $\beta$  are similar. The standard deviations of  $\beta$  are smallest for Poisson and largest for NegBin. The credible intervals based on the quantiles are correspondingly narrowest for Poisson and widest for NegBin. For MixLinkJ8,  $\kappa$  takes on rather large values which effectively reduces  $\text{Var}(Y_i)$  over  $i = 1, \dots, n$ .

Figure 7 plots quantile residuals against predictions and also displays Q-Q plots to assess Normality. The predictions have been computed by taking means of draws from the posterior predictive distribution. Note that there are only 16 distinct values of the covariate  $\mathbf{x}$  and observations with a common covariate are likely to obtain similar predictions. The residuals produced by MixLinkJ8 exhibit the best behavior of the three models, with the least departure from standard Normality. There is still a pattern where smaller predictions tend to have more variable residuals, which indicates that further refinement of the regression function may be needed.

Finally, Figure 8 displays boxplots of  $y$  for each of the 16 possible covariate values, with 95% prediction intervals from both the Poisson and MixLinkJ8 models. These intervals were computed from 2.5% and 97.5% quantiles of the posterior predictive distribution. Intervals for the NegBin model are not shown because the upper limits are far above the range of the plots in all cases. In some cases, the Poisson intervals appear to be too narrow to capture the observed variability of the data, while MixLinkJ8 widens the intervals to reflect the variability.

## 7 Conclusions

Regression on the mean is commonly carried out with exponential family distributions in the Generalized Linear Model framework, but extending this idea to finite mixture distributions is not completely straightforward. This paper formulated the Mixture Link distribution, which establishes a link from a finite mixture

Table 3: DIC for Arizona Medpar models.

Model	DIC
Poisson	2392.62
NegBin	2125.11
MixLinkJ2	2095.07
MixLinkJ3	2096.85
MixLinkJ4	2065.76
MixLinkJ5	2061.04
MixLinkJ6	2062.23
MixLinkJ7	2059.73
MixLinkJ8	2059.39

Table 4: Posterior summaries for Arizona Medpar models.

Poisson	mean	SD	2.5%	97.5%
intercept	1.4947	0.0541	1.3885	1.6012
procedure	0.8447	0.0369	0.7713	0.9161
sex	-0.0292	0.0370	-0.1024	0.0429
admit	0.2813	0.0469	0.1896	0.3749
age75	0.0366	0.0388	-0.0402	0.1092
NegBin	mean	SD	2.5%	97.5%
intercept	1.4972	0.0861	1.3323	1.6698
procedure	0.8492	0.0593	0.7333	0.9634
sex	-0.0422	0.0626	-0.1651	0.0781
admit	0.2889	0.0750	0.1391	0.4366
age75	0.0335	0.0649	-0.0960	0.1628
$\kappa$	0.1938	0.0229	0.1519	0.2416
MixLinkJ8	mean	SD	2.5%	97.5%
intercept	1.5246	0.0759	1.3751	1.6759
procedure	0.9451	0.0507	0.8452	1.0470
sex	-0.0974	0.0526	-0.2013	0.0035
admit	0.2578	0.0627	0.1390	0.3858
age75	0.0849	0.0548	-0.0266	0.1891
$\pi_1$	0.0393	0.0055	0.0280	0.0495
$\pi_2$	0.0631	0.0113	0.0458	0.0931
$\pi_3$	0.1145	0.0158	0.0775	0.1376
$\pi_4$	0.1364	0.0085	0.1181	0.1512
$\pi_5$	0.1472	0.0069	0.1338	0.1609
$\pi_6$	0.1562	0.0071	0.1431	0.1707
$\pi_7$	0.1654	0.0081	0.1515	0.1828
$\pi_8$	0.1779	0.0103	0.1601	0.2008
$\kappa$	17.0029	3.5466	11.0783	24.6940

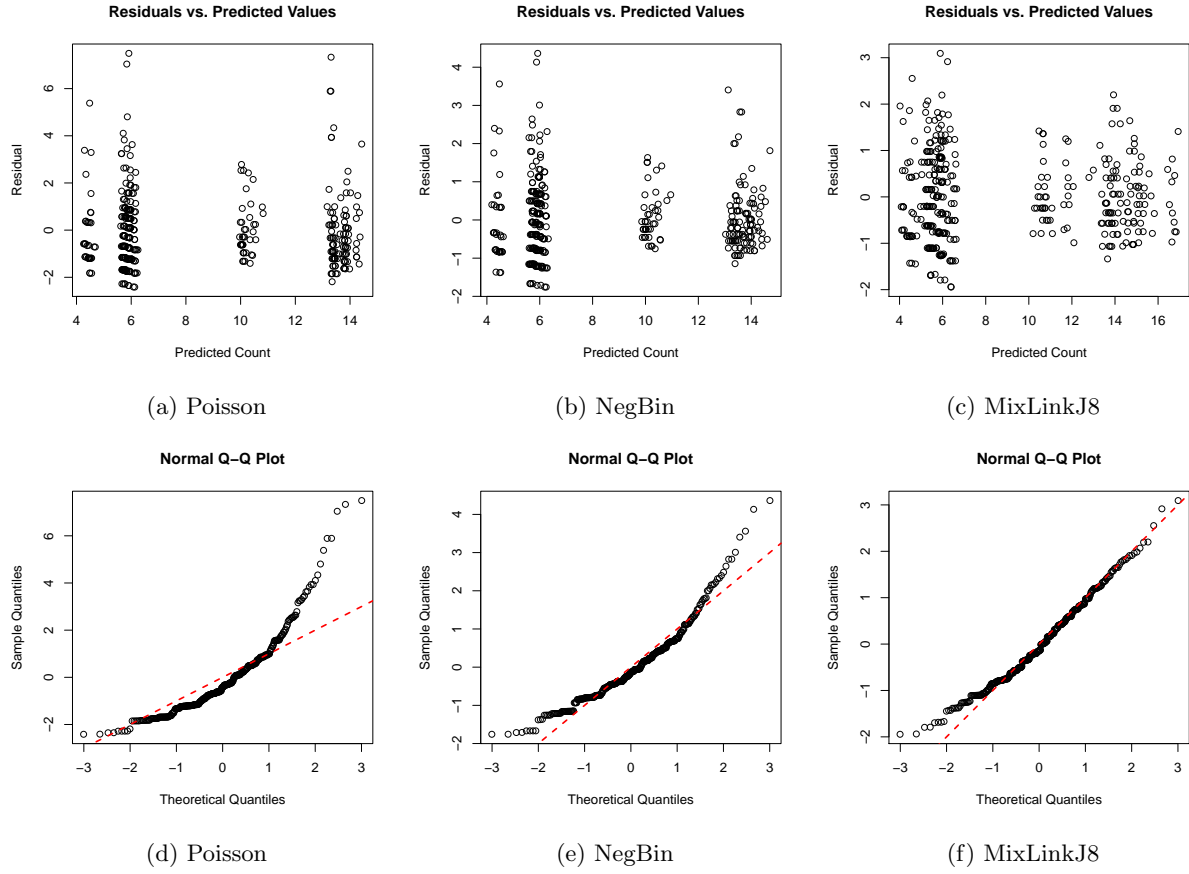


Figure 7: Quantile residuals for Arizona Medpar data.

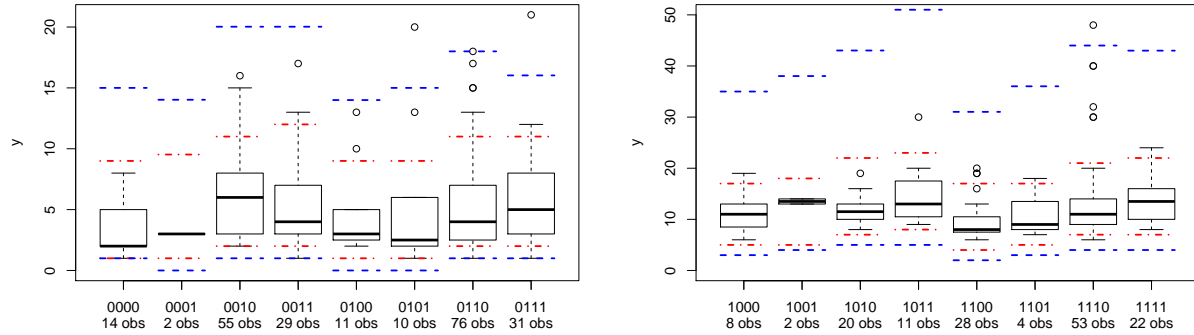


Figure 8: Boxplots of observed  $y_i$  for each of the 16 possible covariate values in the Arizona Medpar data. Covariate values are displayed as a string representing (procedure, sex, admit, age75). For example, “1010” represents procedure = admit = 1 and sex = age75 = 0. Red dash-dot lines represent 95% prediction limits from Poisson and blue dashed lines are from MixLink.

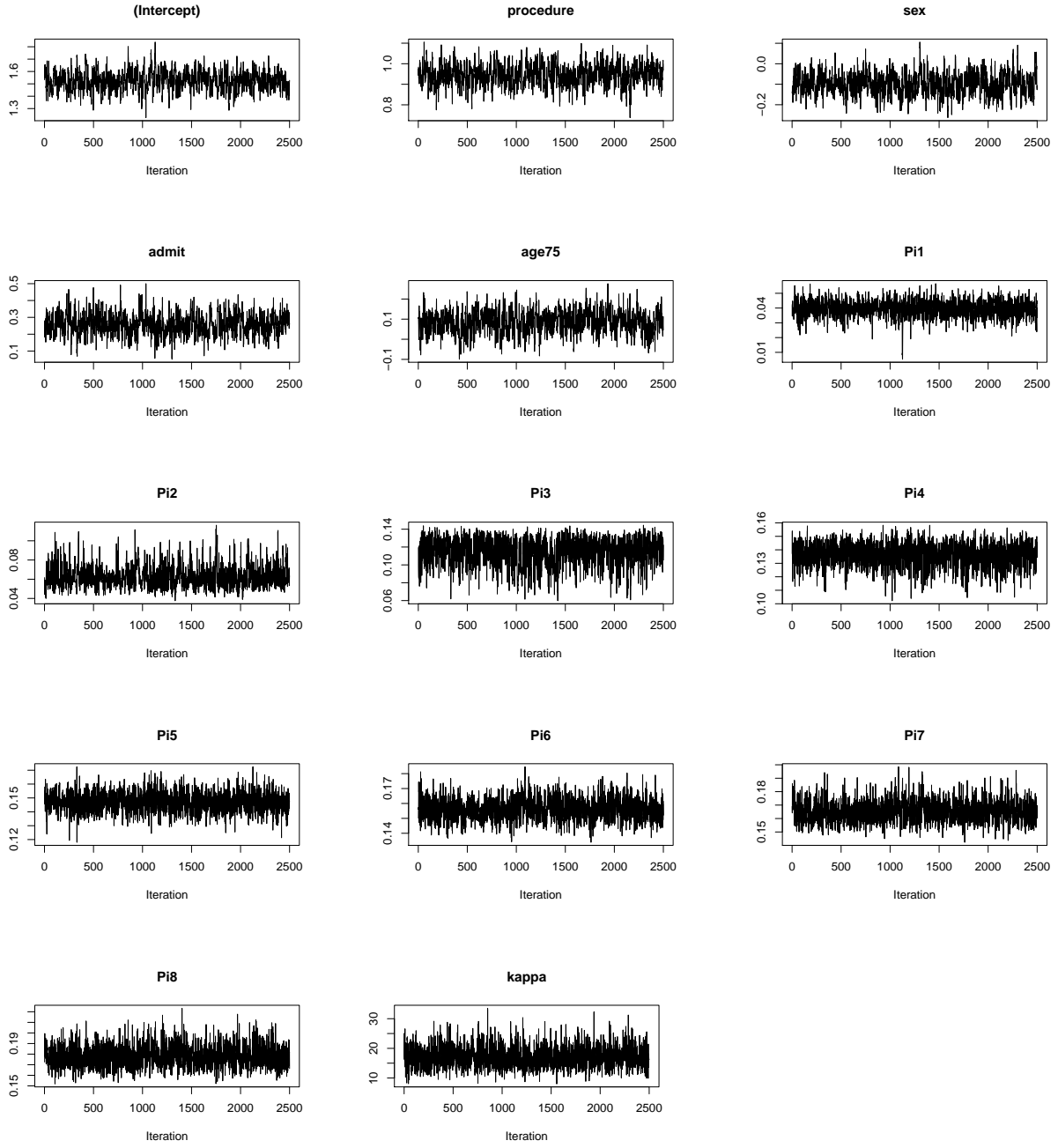


Figure 9: Trace plots for MixLinkJ8 model fit to Arizona Medpar dataset.

mean to the regression function by assuming a random effects structure on the constrained parameter space. Specific variants of Mixture Link were obtained for Binomial, Poisson, and Normal outcomes. Integrals in the general Binomial case appeared not to have a tractable form, but the Normal case could be integrated to yield another (constrained) Normal finite mixture, and integrals in the Poisson case were evaluated using the confluent hypergeometric function. Some interesting connections were noted, for example, between Mixture Link Binomial and the Random-Clumped Binomial and Beta-Binomial distributions. Example regression analyses using Mixture Link Binomial and Poisson models demonstrated utility in handling overdispersion. Simpler models could adequately estimate the regression, yet failed to capture variability seen in the data. This became especially apparent in portions of analysis that depend heavily on the model, such as diagnosing model fit with quantile residuals or computing prediction intervals from the posterior predictive distribution. The fact that Mixture Link is completely likelihood-based ensures that such procedures are available; this could be seen as an advantage over quasi-likelihood methods when a flexible mean-variance relationship is needed. R code for the Mixture Link model is available in the `mixlink` package, available at <http://cran.r-project.org>.<sup>4</sup>

The Mixture Link approach leads to a novel class of distributions with an interesting set of challenges for practical use in data analysis. Initial results in Raim (2014), Raim et al. (2015), and the present paper appear promising, especially using Bayesian inference, but more work is needed to determine the suitability of Mixture Link for wider application. In particular, it may be worthwhile to investigate analytical properties of Mixture Link models, such as differentiability, especially in the Binomial case. Such properties may be needed to establish appropriate methods for maximum likelihood estimation, large sample properties of maximum likelihood estimates, and approximation of the posterior distribution by a Normal distribution.

## Acknowledgements

We thank Professors Thomas Mathew, Yi Huang, and Yaakov Malinovsky at the University of Maryland, Baltimore County (UMBC) for serving on the committee of the dissertation in which this work was initiated. We thank the UMBC High Performance Computing Facility for use of its computational resources, and for financial support of the first author through a multiple year graduate assistantship.

## A Appendix: MCMC for Binomial Mixture Link

An MCMC algorithm based on model (3.8) can be formulated with  $\psi_{ij}$  as augmented data. This approach avoids expensive numerical integration needed to compute the likelihood. The joint distribution of all random quantities is

$$f(\mathbf{y}, \boldsymbol{\psi}, \boldsymbol{\beta}, \boldsymbol{\pi}, \kappa) = \left\{ \prod_{i=1}^n Q(y_i, \boldsymbol{\psi}_i, \boldsymbol{\beta}, \boldsymbol{\pi}, \kappa) \right\} f(\boldsymbol{\beta}) f(\boldsymbol{\pi}) f(\kappa),$$

$$\text{where } Q(y_i, \boldsymbol{\psi}_i, \boldsymbol{\beta}, \boldsymbol{\pi}, \kappa) = \sum_{j=1}^J \pi_j \text{Bin}(y_i \mid m_i, H_{ij}(\psi_{ij})) \mathcal{B}(\psi_{ij} \mid a_{ij}, b_{ij}),$$

and  $H_{ij}(x) = (u_{ij} - \ell_{ij})x + \ell_{ij}$ . Gibbs steps to sample  $\boldsymbol{\beta}$ ,  $\boldsymbol{\pi}$ ,  $\kappa$ , and  $\boldsymbol{\Psi} = \{\boldsymbol{\psi}_i : i = 1, \dots, n\}$  will not yield closed forms. Instead, we will use simple Random Walk Metropolis Hastings (Robert and Casella, 2010, Section 7.5) to propose draws for each random quantity.

To obtain draws of the constrained parameters  $\boldsymbol{\pi}$ ,  $\kappa$ , and  $\boldsymbol{\Psi}$ , we draw unconstrained random variables from the sampler and transform them to the constrained space. Generally, denote  $\boldsymbol{\xi}$  as one of the constrained parameters whose full conditional density is  $f(\boldsymbol{\xi} \mid \text{Rest})$ , and let  $h$  be a bijection from the space of  $\boldsymbol{\xi}$  to a

---

<sup>4</sup>The package currently provides Mixture Link Binomial and Poisson distributions and MCMC samplers. Functions to compute maximum likelihood estimates using numerical optimization are also implemented.

Euclidean space  $\mathbb{R}^k$ . The density of  $\phi = h(\xi)$  is then  $f(h^{-1}(\phi) \mid \text{Rest}) |\det \mathfrak{J}(\phi)|$ , where  $\mathfrak{J}(\phi) = \partial \xi / \partial \phi$ . Starting from a given  $\phi = h(\xi)$ , a proposed  $\phi^*$  will be accepted with probability

$$\min \left\{ 1, \frac{f(h^{-1}(\phi^*) \mid \text{Rest}) \cdot |\det \mathfrak{J}(\phi^*)|}{f(h^{-1}(\phi) \mid \text{Rest}) \cdot |\det \mathfrak{J}(\phi)|} \right\}.$$

Note that the function  $Q(y_i, \psi_i, \beta, \pi, \kappa)$  needs to be evaluated in each step. By computing  $Q$  in `C/C++`, it is possible to improve the performance greatly over a pure `R` ([R Core Team, 2015](#)) implementation of our sampler. The `Rcpp` package by [Eddelbuettel and Francois \(2011\)](#), for example, greatly facilitates a hybrid implementation of `R` and `C++`.

**Gibbs step for  $\beta$ .** Consider the unnormalized density

$$q(\beta \mid \text{Rest}) = \left\{ \prod_{i=1}^n Q(y_i, \psi_i, \beta, \pi, \kappa) \right\} f(\beta).$$

Suppose  $\beta^{(r)}$  is the current iterate of  $\beta$  in the simulation and draw  $\beta^*$  from the proposal distribution  $N(\beta^{(r)}, \mathbf{V}_\beta^{\text{prop}})$ . Draw  $U \sim \mathcal{U}(0, 1)$ , and let

$$\beta^{(r+1)} = \begin{cases} \beta^* & \text{if } U < \frac{q(\beta^* \mid \text{Rest})}{q(\beta^{(r)} \mid \text{Rest})} \\ \beta^{(r)} & \text{otherwise.} \end{cases}$$

**Gibbs step for  $\pi$ .** Consider the unnormalized density

$$q(\pi \mid \text{Rest}) = \left\{ \prod_{i=1}^n Q(y_i, \psi_i, \beta, \pi, \kappa) \right\} f(\pi).$$

Suppose  $\pi^{(r)}$  is the current iterate of  $\pi$  in the simulation. Denote  $\mathbb{S}^J$  as the probability simplex in dimension  $J$  with typical element  $\mathbf{p} = (p_1, \dots, p_J)$ . Note that the multinomial logit function  $h(\mathbf{p}) = (\log(p_1/p_J), \dots, \log(p_{J-1}/p_J))$  is a bijection from  $\mathbb{S}^J$  to  $\mathbb{R}^{J-1}$ . Therefore, we can draw  $\phi^*$  from the proposal distribution  $N(h(\pi^{(r)}), \mathbf{V}_\pi^{\text{prop}})$  on  $\mathbb{R}^{J-1}$  and let  $\pi^* = h^{-1}(\phi^*)$  be the candidate for the next iterate. Denote  $\mathfrak{J}(\phi) = \frac{\partial \pi}{\partial \phi}$  as the  $J \times (J-1)$  Jacobian of the transformation from  $\phi$  to  $\pi$ , and let  $\det \mathfrak{J}(\phi)$  be the determinant ignoring the  $J$ th row. Draw  $U \sim \mathcal{U}(0, 1)$ , and let

$$\pi^{(r+1)} = \begin{cases} \pi^* & \text{if } U < \frac{q(\pi^* \mid \text{Rest})}{q(\pi^{(r)} \mid \text{Rest})} \frac{|\det \mathfrak{J}(\phi^*)|}{|\det \mathfrak{J}(\phi^{(r)})|} \\ \pi^{(r)} & \text{otherwise.} \end{cases}$$

**Gibbs step for  $\kappa$ .** Consider the unnormalized density

$$q(\kappa \mid \text{Rest}) = \left\{ \prod_{i=1}^n Q(y_i, \psi_i, \beta, \pi, \kappa) \right\} f(\kappa).$$

Suppose  $\kappa^{(r)}$  is the current iterate of  $\kappa$  in the simulation. Draw  $\phi^*$  from the proposal distribution  $N(\log(\kappa^{(r)}), V_\kappa^{\text{prop}})$  and let  $\kappa^* = \exp(\phi^*)$  be the candidate for the next iterate. The Jacobian of the transformation from  $\phi$  to  $\kappa$  is  $\frac{\partial \kappa}{\partial \phi} = \exp(\phi)$ . Draw  $U \sim \mathcal{U}(0, 1)$ , and let

$$\kappa^{(r+1)} = \begin{cases} \kappa^* & \text{if } U < \frac{q(\kappa^* \mid \text{Rest})}{q(\kappa^{(r)} \mid \text{Rest})} \frac{\exp(\phi^*)}{\exp(\phi^{(r)})} \\ \kappa^{(r)} & \text{otherwise.} \end{cases}$$

**Gibbs step for  $\psi$ .** Consider the unnormalized density

$$q(\psi \mid \text{Rest}) = \prod_{i=1}^n Q(y_i, \psi_i, \beta, \pi, \kappa).$$

We can see that  $\psi_i$  are independent conditional on the remaining random variables and we may therefore consider drawing one at a time. Suppose  $\psi_i^{(r)}$  is the current iterate of  $\psi_i$  in the simulation. Let  $G$  be the CDF of the logistic distribution, which is a bijection from  $\mathbb{R}$  to the unit interval. Denote  $\phi^{(r)} = (G^{-1}(\psi_{i1}^{(r)}), \dots, G^{-1}(\psi_{iJ}^{(r)}))$ . The Jacobian of the transformation from  $\phi$  to  $\psi_i$  is

$$\frac{\partial \psi_i}{\partial \phi} = \text{Diag}(G'(\phi_1), \dots, G'(\phi_J)) \implies \det \left( \frac{\partial \psi_i}{\partial \phi} \right) = \prod_{j=1}^J G'(\phi_j),$$

where  $G'$  represents the logistic density. Draw  $\phi^*$  from the proposal distribution  $N(\phi^{(r)}, V_\phi^{\text{prop}})$  and let  $\psi_i^* = (G(\phi_1^*), \dots, G(\phi_J^*))$  be the candidate for the next iterate. Draw  $U \sim \mathcal{U}(0, 1)$ , and let

$$\psi_i^{(r+1)} = \begin{cases} \psi_i^* & \text{if } U < \frac{q(\psi_i^* \mid \text{Rest})}{q(\psi_i^{(r)} \mid \text{Rest})} \frac{\prod_{j=1}^J G'(\phi_j^*)}{\prod_{j=1}^J G'(\phi_j^{(r)})}, \\ \psi_i^{(r)} & \text{otherwise.} \end{cases}$$

## References

- Alan Agresti. *Categorical Data Analysis*. Wiley-Interscience, 2nd edition, 2002.
- Akio A. Awa, Takeo Honda, Toshio Sofuni, Shotaro Neriishi, Michihiro C. Yoshida, and Takashi Matsui. Chromosome-aberration frequency in cultured blood-cells in relation to radiation dose of A-bomb survivor. *The Lancet*, 298(7730):903–905, 1971.
- Sanjib Basu and Saurabh Mukhopadhyay. Binary response regression with normal scale mixture links. In Bani K. Mallick Dipak K. Dey, Sujit K. Ghosh, editor, *Generalized Linear Models: A Bayesian Perspective*, pages 231–242. CRC Press, 2000.
- Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.
- Michelle R. Danaher, Anindya Roy, Zhen Chen, Sunni L. Mumford, and Enrique F. Schisterman. Minkowski-Weyl priors for models with parameter constraints: An analysis of the biocycle study. *Journal of the American Statistical Association*, 107(500):1395–1409, 2012.
- Dipak K. Dey and Nalini Ravishanker. Bayesian approaches for overdispersion in generalized linear models. In Bani K. Mallick Dipak K. Dey, Sujit K. Ghosh, editor, *Generalized Linear Models: A Bayesian Perspective*, pages 73–88. CRC Press, 2000.
- Dipak K Dey, Sujit K Ghosh, and Bani K Mallick. *Generalized linear models: A Bayesian perspective*. CRC Press, 2000.
- Peter K. Dunn and Gordon K. Smyth. Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, 5(3):236–244, 1996.
- Dirk Eddelbuettel and Romain Francois. Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, 40(1):1–18, 2011.
- Sylvia Frühwirth-Schnatter. *Finite Mixture and Markov Switching Models*. Springer, 2006.

- Daniel B. Hall. Zero-inflated poisson and binomial regression with random effects: A case study. *Biometrics*, 56(4):1030–1039, 2000.
- James W. Hardin and Joseph M. Hilbe. *Generalized Estimating Equations*. Chapman and Hall/CRC, 2nd edition, 2012.
- Joseph M. Hilbe. *Negative Binomial Regression*. Cambridge University Press, 2nd edition, 2011.
- A. Jasra, C. C. Holmes, and D. A. Stephens. Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statistical Science*, 20(1):50–67, 2005.
- Norman L. Johnson, Samuel Kotz, and Adrienne W. Kemp. *Univariate Discrete Distributions*. Wiley-Interscience, 3rd edition, 2005.
- Nadja Klein, Thomas Kneib, and Stefan Lang. Bayesian generalized additive models for location, scale, and shape for zero-inflated and overdispersed count data. *Journal of the American Statistical Association*, 110(509):405–419, 2015.
- Chuanhai Liu and Donald B. Rubin. ML estimation of the t distribution using EM and its extensions, ECM and ECME. *Statistica Sinica*, 5:19–39, 1995.
- P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman and Hall/CRC, 2nd edition, 1989.
- Charles E. McCulloch, Shayle R. Searle, and John M. Neuhaus. *Generalized, Linear, and Mixed Models*, volume 2. Wiley-Interscience, 2nd edition, 2008.
- Jorge G. Morel and Neerchal K. Nagaraaj. A finite mixture distribution for modelling multinomial extra variation. *Biometrika*, 80(2):363–371, 1993.
- Jorge G. Morel and Nagaraaj K. Neerchal. *Overdispersion Models in SAS*. SAS Institute, 2012.
- Masanori Otake and Ross L. Prentice. The analysis of chromosomally aberrant cells based on beta-binomial distribution. *Radiation research*, 98(3):456–470, 1984.
- Serge B. Provost and Young-Ho Cheong. On the distribution of linear combinations of the components of a dirichlet random vector. *Canadian Journal of Statistics*, 28(2):417–425, 2000.
- R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015.
- Andrew M. Raim. Computational methods in finite mixtures using approximate information and regression linked to the mixture mean. Ph.D. Thesis, Department of Mathematics and Statistics, University of Maryland, Baltimore County, 2014.
- Andrew M. Raim, Marissa N. Gargano, Nagaraaj K. Neerchal, and Jorge G. Morel. Bayesian analysis of overdispersed binomial data using mixture link regression. In *JSM Proceedings, Statistical Computing Section*. Alexandria, VA: American Statistical Association, pages 2794–2808, 2015.
- Christian P. Robert and George Casella. *Monte Carlo Statistical Methods*. Springer, 2nd edition, 2010.
- T. Sofuni, T. Honda, M. Itoh, S. Neriishi, and M. Otake. Relationship between the radiation dose and chromosome aberrations in atomic bomb survivors of Hiroshima and Nagasaki. *Journal of Radiation Research*, 19(2):126–140, 1978.
- Martin A. Tanner and Wing Hung Wong. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82(398):528–540, 1987.
- R. W. M. Wedderburn. Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika*, 61(3):439–447, 1974.